THE SIMULATOR SAYS YES: LANGUAGE WORLD MODELS NARRATE THE EXECUTIONS THEIR ENVIRONMENT REFUSED

Anonymous authors
Paper under double-blind review

ABSTRACT

Language world models—models fine-tuned on agent trajectories to predict the next tool observation—are now used as environments for training agents with reinforcement learning. We show that they systematically answer the wrong question. An environment must say what an action does and whether it is allowed; trajectories supervise the first on every step and give few examples of the second (1.5% of rows in the corpus we study). On 399 held-out actions the real executor refused, world models fine-tuned from Qwen3 (1.7B–8B) report a normal outcome instead of the refusal on 57 / 44 / 32% and the released 35B Qwen-AgentWorld on 25%—a failure we call *whitewash*—and the pretrained checkpoints, prompted, whitewash as much or more: fine-tuning removes only part of a deficit that predates it. Even on the rows an LLM reader (checked by one author on 100) judges decidable from the shown context, the 1.7B model still whitewashes 0.54 and the released model 0.14: much of the deficit is not missing information but a failure to use it. More refusal data helps, but not for the right reason: raising the 1.7B's refusal share from 1.5% to 21.3% lowers its whitewash from 0.57 to 0.22 (0.54 to 0.16 on the decidable rows, level with the released 35B) at a false-refusal cost, yet leaves a linear probe of the state unchanged and barely reaches planted state-dependent violations; in a τ-bench corpus where every refusal follows a missing read, the simulator learns that cue and no share moves it, which a two-prefix test exposes. Trained inside a simulator that never refuses, a GRPO policy stops refusing within twenty steps, because the task reward makes that optimal; the executor's refusals, passed through per call or used to threshold a learned head, restore it. What intermediate whitewash costs a policy is not established here. We release the refusal set, the planted battery and the instrument.
Code: https://anonymous.4open.science/r/whitewash-audit-6635

1 INTRODUCTION

A language model fine-tuned on agent trajectories to predict the next tool observation becomes, once a policy is trained inside it, an environment (Chen et al., 2026; Zuo et al., 2026; Deshpande, 2026; Patronus AI, 2026). An environment answers two questions about an action: what happens, and whether it is allowed. The real executor answers both: it runs the action and refuses when the tool does not exist, the arguments are malformed, or the file is missing. The trajectories of competent agents answer the first question on every row and the second almost never: in the 30,000-row training slice we use, 444 rows (1.5%) begin with an executor error, and the pipeline behind the released Qwen-AgentWorld models is described as removing error-and-retry turns before training (Zuo et al., 2026).

This matters because a policy trained inside a simulator is rewarded by what the simulator says happened. We call the failure **whitewash**: the simulator reports no refusal at all where the executor refused. Figure 1 shows it on one held-out row, how often it happens, and that raising a corpus's refusal share lowers it in one corpus and not in another whose refusals all follow one cue. This paper measures whitewash and where it sits (§3), what moves it and at what price (§4), and what restores a policy's refusals when its simulator never emits any (§5). That a simulator sometimes misses an error

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

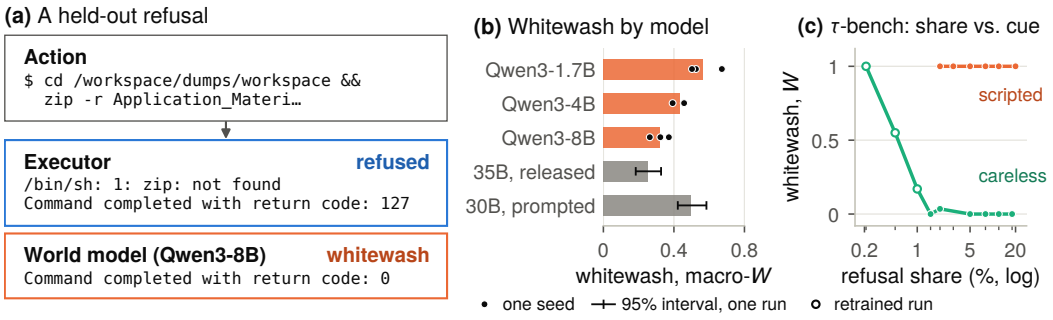


Figure 1: **(a)** One held-out refusal: the executor refuses the action, the Qwen3-8B world model predicts that it ran. **(b)** Whitewash on the 399 held-out real refusals: the three Qwen3 world models we fine-tune (orange bars at the seed mean, one dot per seed) and two we did not train (gray; 95% intervals; the released model read in its own format). **(c)** Whitewash against the training corpus's refusal share in τ-bench retail (one run per share, a tick per rung, retrained runs hollow): no share moves the simulator whose refusals all follow a missing read; the corpus rebuilt so that they follow a read reports them from 1.5% on (§4.3).

is documented (Qwen-AgentWorld's error-handling axis scores its text domains 60–69%; Zuo et al., 2026); which refusals, and what moves them at what price, is not.

Our contributions:

1. **An admissibility audit on executor-refused rows, and its artefacts (§3).** On 399 held-out refusals, Qwen3 world models at 1.7B / 4B / 8B whitewash 0.57 / 0.44 / 0.32 (three seeds each; 0.54 / 0.40 / 0.24 on the rows an LLM reader, checked by one author on 100 rows, calls decidable from the prefix), the released 35B model 0.25 (0.14), and the pretrained checkpoints as much or more; on the two state-decided classes of a planted battery (validated by recipe, not executed) no unrepaired model recovers more than 8% of the executor's refusals.
2. **The lever, its price and its limit (§4).** Raising the refusal share of the 1.7B training mix from 1.5% to 21.3% moves whitewash from 0.57 to 0.22 while false refusals rise from 0.05 to 0.09 (three seeds at each end, one run between); a linear probe of the state does not move, and the gain barely reaches the battery. In a τ-bench corpus we construct so that every refused row follows a missing read, the simulator refuses iff nothing was read and no share moves it; a two-prefix test exposes the cue.
3. **A sanity check in a live executor, and what restores the behaviour (§5).** In a τ-bench simulator with $W = 1.000$, where the task reward makes never refusing optimal by construction, GRPO reaches that optimum by step 20. Supplying the executor's refusals restores it under this reward, per call (1.000 on every seed) or as executor-labelled goals that threshold a learned head (0.972); under an earlier rule reward the same pass-through collapsed the policy (Table E.14). The one non-degenerate observation is a simulator at $W = 0.035$, whose policy keeps the refusal behaviour on two of three seeds (Table E.16).

Scope. One corpus supplies the natural rows and the lever is measured at 1.7B; the battery's state classes are SWE-only; the cue is identified only in the constructed corpus; and no simulator with the intermediate whitewash of §3 was used as an environment, so what that whitewash costs a policy is not established here. Throughout, validity factors as $V(s, z) = V_{\text{syntax}}(z) \wedge V_{\text{schema}}(z) \wedge V_{\text{state}}(s, z)$; the first two conjuncts are decidable by static checks, the third only from the state. Every reading was registered in a dated file before its data were read; wherever the paper reports a reading other than the registered one, Table G.1 prints both, and the two analyses registered after their data are labelled post hoc (Appendix G).

2 Setup

Corpus, models, training. We LoRA-fine-tune Qwen3-1.7B, 4B, and 8B, and SmolLM2-1.7B as a second family, on a 30,000-row slice of C10v2, the Patronus world-model corpus (SWE editing and

Table 1: Whitewash W on the 399 natural refusals (macro over the four classes), on the 230 of them an LLM reader, checked by an author on 100 rows, calls decidable from the prefix (macro over the three qualifying classes; Appendix C.6), and false refusals on the 185 tool-matched legal companions (head $\vee$ Y). Parentheses: seed mean with the three-seed range; brackets: one run with its 95% template-clustered interval; the repaired arm's decidable value is its first seed; SmolLM2 and the released model are read on the 361 and 393 rows their samples tag (Appendix C.1).

model	training corpus, refusal share	W, all 399 rows	W, 230 decidable rows	false refusals (legal)
Qwen3-1.7B	C10v2 slice, 1.5%	0.566 (0.503–0.672)	0.544 (0.470–0.650)	0.050 (0.034–0.067)
Qwen3-4B	same	0.435 (0.393–0.458)	0.400 (0.366–0.435)	0.042 (0.040–0.045)
Qwen3-8B	same	0.319 (0.263–0.371)	0.239 (0.171–0.274)	0.036 (0.025–0.050)
SmolLM2-1.7B	same	0.676 [0.600, 0.743]	0.643 [0.568, 0.731]	0.064 [0.027, 0.107]
Qwen-AgentWorld-35B-A3B	released	0.253 [0.183, 0.327]	0.137 [0.090, 0.200]	0.031 [0.007, 0.064]
Qwen3-30B-A3B-Instruct	none (prompted)	0.497 [0.422, 0.584]	0.359 [0.275, 0.469]	0.031
1.7B + 724 verbatim refusals	slice + bank, 21.3%	0.217 (0.181–0.237)	0.156 [0.108, 0.202]	0.088 (0.070–0.110)
1.7B, own 444 upsampled	own rows, 21.3%	0.208 [0.142, 0.283]	—	0.119

execution, API tools; Patronus AI, 2026), to predict the next observation (one epoch; Reproducibility Statement). Each model must pass an admission check (canonical-direction accuracy ≥ 0.80 on held-out legal actions) before we read the battery; the released Qwen-AgentWorld-35B-A3B (Zuo et al., 2026) scores 0.774 under our prompt formats, so we read it only on the natural rows, in its own `<predicted_observation>` format (Appendix C.4).

The natural refusal set. Scanning the corpus for recorded observations that begin with an executor error line finds 3,251 verified refusals in the train split (1.36%) and 700 more in the held-out splits; from the latter we froze 399 refused rows and 185 tool-matched legal companions before any model ran. Rows are capped at ten per refusal template; weighting templates by census frequency instead lowers or leaves W for every arm (by 0.03–0.04 for the three Qwen3 world models) and changes no ordering (Appendix A.2). The corpus's tools are a mix of live code execution, recorded retrievals and deterministic mock services (32 / 42 / 325 of the 399; none is a language model's output, Appendix A.2). Each refusal falls in one of four classes: **nonexistent-entity** and **unmet-precondition**, which only the state decides, and **unavailable-tool** and **invalid-args**, which a static schema check could catch (at most 129 rows, 32%).

The planted battery. On 200 recorded states from the test split, none in any training slice, we plant executor-invalid variants of recorded actions that are parseable, well-formed, and refused: planted entities and edit targets are random tokens checked absent from the state, planted tool names ones the state never offers (Appendix A.3). The four recipes, instantiated on the tools of a live executor (τ-bench retail; Yao et al., 2024), error 240/240 times across all 37 of its rejection paths; this validates the recipes rather than the individual items, whose states we cannot re-run.

Instruments. An emission *reports a refusal* if it opens with an executor error line (the *head rule*), reproduces the template of the recorded refusal, or is flagged by Y, a frozen lexical error classifier (Appendix B.1, Figure A.1). On the natural rows we credit any of the three, so the instrument is deliberately permissive, and whitewash is $W = 1 - R$, macro-averaged over the four classes with a template-clustered bootstrap ($B = 2000$). The companions drop the template rule, which on a legal row detects fidelity rather than refusal, and read false refusals with head $\vee$ Y; the battery uses the head rule alone, since Y fires on file listings. A wrapper-agnostic detector moves W by at most 0.002 on any model we train (Appendix B.3).

3 Learned simulators whitewash refusals

3.1 Natural refusals, out of sample

Table 1 reads every model at three samples per row, on all 399 rows and on the 230 rows an LLM reader, checked below, calls decidable from the shown prefix. Whitewash falls with scale and does not vanish: within Qwen3 the three seed ranges are pairwise disjoint on all rows (0.57 / 0.44 / 0.32), and the decidable rows read 0.54 / 0.40 / 0.24. Most of the whitewash is success narration: of the 8B's 1,194 samples on the refused rows, 0.21 [0.16, 0.26] narrate a success or a normal observation where the executor refused (0.83 of that run's pooled whitewash), against 0.01 that are the executor's own refusal missed by the instrument (Appendix C.5).

Table 2: Rate at which the emission begins with an executor error line on the planted battery, with the matched legal baseline (the same model on the same states' legal action) in parentheses; state-clustered 95% CIs in Table C.3. Matched macro-BR: boundary recovery, the fraction of the distance from the model's rate on the legal baseline to the executor's certainty, averaged over classes. n: states per class, Qwen · SmolLM2.

class	n	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	SmolLM2-1.7B
nonexistent-entity (state)	$44 \cdot 26$	0.000 (0.000)	0.009 (0.000)	0.050 (0.023)	0.000 (0.000)
unmet-precondition (state)	$45 \cdot 26$	0.000 (0.000)	0.040 (0.004)	0.080 (0.040)	0.000 (0.000)
unavailable-tool (static)	$199 \cdot 144$	0.035 (0.008)	0.018 (0.009)	0.015 (0.017)	0.025 (0.011)
invalid-args (static)	$45 \cdot 27$	0.004 (0.000)	0.049 (0.004)	0.222 (0.040)	0.000 (0.000)
matched macro-BR		0.008	0.025	0.064	0.004
95% CI		[0.00, 0.01]	[0.01, 0.04]	[0.04, 0.09]	[0.00, 0.01]

The decidable rows come from an LLM read of all 399, checked against an author's read of 100 (agreement 0.85, κ 0.69), which finds the fact that decides the refusal in the shown prefix on 230 rows; both reads saw the recorded refusal, so they grade explicability after the fact, and the reader's bias runs toward decidable (Appendix C.6). On those rows the 1.7B model whitewashes 0.54 [0.48, 0.62], against 0.66 [0.57, 0.75] on the 168 undecidable rows (the released model 0.14 against 0.48). Undecidable rows have no zero floor (a simulator calibrated to the corpus would whitewash them at $1 - P(\text{refusal} \mid \text{prefix})$), so they describe the corpus's prefixes as much as the model; 0.54 is the estimate on the rows labelled decidable, not a bound, and the ordering holds on one class set for every stratum and on the rows both readers call decidable (0.48 at 1.7B, 0.09 released; Table C.9). The registered 1.7B contrast, $+0.12$ [$+0.00$, $+0.22$], is marginal and carried by invalid-args ($+0.43$; the two state classes $+0.03$ and -0.10), and a withdrawn earlier read would have placed it at $+0.36$; at 4B, 8B, the released and the prompted arms the contrast is larger, $+0.17$ to $+0.34$ with intervals excluding zero (Table C.7).

The released flagship, Qwen-AgentWorld-35B-A3B, asked in its own format, whitewashes 0.25 [0.18, 0.33] of the 399 rows and 0.14 of the decidable ones. On 48 refusals from its own benchmark it reports no refusal on 0.39 [0.26, 0.53] and narrates a success on 0.33 (two human readers; Appendix C.4). Qwen3-30B-A3B-Instruct, prompted in the flagship's format with no training, whitewashes 0.50 [0.42, 0.58].

The deficit predates fine-tuning. The Qwen3 base checkpoints, prompted under the flagship's format (a different contract from the fine-tuned arms), whitewash 0.68 / 0.44 / 0.41 at 1.7B / 4B / 8B once the refusals they report inside their own JSON error wrapper are counted (Appendix C.3). Paired row by row, fine-tuning on the 1.5% corpus lowers W by 0.14 at 1.7B and 0.18 at 8B and leaves 4B unchanged ($+0.03$ [-0.03, $+0.09$]; Appendix C.3).

Two human readers checked the instrument on a 220-row sheet (20–40 emissions per arm, identities hidden): they agree with each other on the 217 rows both labelled ($\kappa = 0.93$) and with the instrument at $\kappa \geq 0.60$ on nine of ten arms, the exception being the Qwen3-4B base checkpoint, whose JSON-wrapped refusals the frozen rules miss (every disagreement is printed in Appendix B.5).

3.2 The planted battery: the state classes are near zero

Table 2 reads the battery with the head rule, five samples per state, beside each model's false-refusal rate on the same states' canonical legal action. Only the static invalid-args class exceeds a boundary recovery of 0.10, and only at 8B under the head rule (0.19 [0.11, 0.29]; 0.17 at 4B and 8B under Y). On the two state classes no unrepaired model's matched recovery exceeds 0.042 under the head rule or 0.078 under Y. The typical emission is a normal observation: on nonexistent-entity (an editor view of a path that does not exist), 89–95% of every model's emissions are a fabricated `cat -n` listing of the missing file.

The difference between Table 2 and Table 1 comes from the rows rather than from the instrument: under the head rule alone the same models report 0.38 / 0.46 / 0.66 of the natural refusals (Table B.3) and at most 0.08 of the battery's state-class violations (Table 2). Part is domain: the battery's state classes sit on SWE states, and natural SWE rows whitewash more than API rows (Appendix C.1).

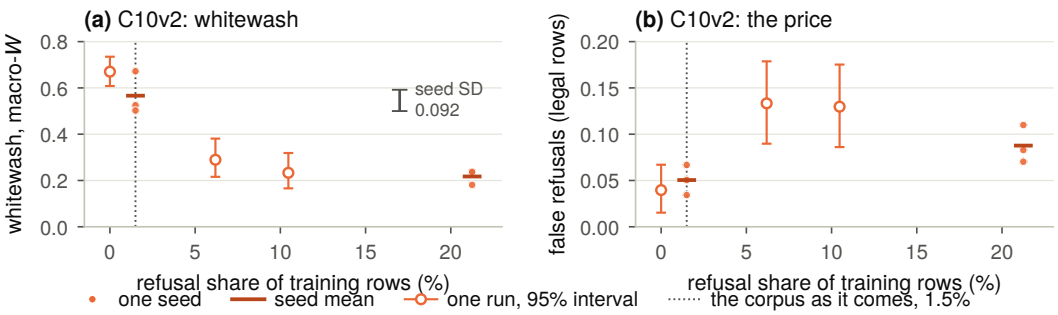


Figure 2: Whitewash and its price against the refusal share of the C10v2 training mixture. **(a)** Macro-W on the 399 natural refusals: at 1.5% and 21.3% one dot per seed with a bar at the seed mean; at 0, 6.2 and 10.5% one run, hollow, with its 95% template-clustered interval (a row-level interval, which carries no run-to-run variance); the scale bar is the recipe's seed SD at 1.7B. **(b)** The same runs, scored for false refusals on the 185 tool-matched legal companions.

4 Refusal share moves what the simulator says; the corpus's cue decides whether it can

4.1 Refusal share sets what the model says

The repair. Direct supervision for a refused action is rare in C10v2: 444 of the 30,000 training rows (1.5%) begin with an executor error. The obvious repair fails first: continued SFT of the 1.7B model on refusal targets harvested by the lexical classifier Y does not lower W (0.60 and 0.61 at two doses, single runs, inside the reference's seed range 0.50–0.67), because most harvested targets were `cat -n` listings Y had mislabelled as errors. The repair that works keeps the recipe and replaces the *bank* (the refusal rows added to the training mix) with the train split's own executor refusals, verbatim: 724 head-rule rows of the two static classes, repeated to a 21.3% share of the training mix. It reaches $W = 0.22$ over three seeds (Welch t: -0.35 [-0.55, -0.15]).

The ladder. At five shares (*rungs*) of the same recipe (0 / 1.5 / 6.2 / 10.5 / 21.3%), W falls 0.67 / 0.57 / 0.29 / 0.23 / 0.22, with three seeds at 1.5% and 21.3% and one run at 0, 6.2 and 10.5% (Figure 2a). The endpoints separate far outside the recipe's seed SD (0.092 at 1.7B); two of the four interior steps are smaller than it, so we report the interior ordering without claiming it. False refusals on legal rows run 0.04–0.13 across the shares and peak at the single-run interior rungs (Figure 2b); the fall is not bought by blanket refusal.

What it costs. In seed means at the ladder's ends, false refusals rise from 0.05 to 0.09 while whitewash falls from 0.57 to 0.22. At the corpus's own 1.5% refusal share that is about seven new false refusals per refusal recovered, at equal cost per error; the repair breaks even where about 10% of actions are refused. The figure is illustrative for this sample's composition; what a false refusal costs a policy in RL we do not measure.

Bank or share? A 2×2 that crosses the bank (the corpus's own 444 rows, or a new 724-row bank) with the share (1.5% or 21.3%) reads share. Read as single runs (Table D.3), raising the share moves W from 0.52 to 0.21 with the corpus's own rows and from 0.57 to 0.18 with the new bank; at 1.5% both banks have three seeds and their seed sets overlap (Appendix D.2). We cannot distinguish upsampling the refusals the corpus already holds from adding new ones (0.21 against 0.22, the former at a false-refusal rate of 0.12, one run; Table 1, last two rows).

Reach. The repair is corpus-local: W falls $0.40 \to 0.07$ on the 99 natural rows the bank's template groups cover and $0.56 \to 0.21$ on the 288 they never contained; on the battery, whose tools the bank excludes, matched recovery over the four classes rises from 0.008 to 0.064 and no further; the bank holds only the two static classes, so this is transfer across classes, and the own-rows arm was not read on the battery. It also lowers W as much on rows whose prefix does not decide the refusal ($0.66 \to 0.27$) as on rows it does ($0.54 \to 0.16$; Table C.7): the share raises how often the corpus's

Table 3: Share of the simulator's predictions that are a refusal, 200 write calls per cell, by whether the executor allows or refuses the call (columns) and whether the prefix contains a read of the order (rows); one run each. Left: the simulator of §5.3, whose refused training rows all follow a missing read. Right: the same recipe rebuilt from a careless agent that reads first and still writes illegally.

	scripted-corpus simulator (2%)		careless-corpus simulator (2%)	
prefix of the write call	legal	executor-refused	legal	executor-refused
the agent has read the order	0.000	0.000	0.000	0.990
no read in the prefix	1.000	1.000	0.095	0.055

refusals are emitted on the corpus's cue; that it teaches inference from the state, neither the probe nor the battery shows.

4.2 Not rarity alone, and not the representation

Rarity. A comparably rare *legal* observation class (2.22% of rows) is reproduced on a retrained 1.7B checkpoint (under-reproduction $U = 0.04$ [0.02, 0.05]) where the refusals are not ($W = 0.63$): rarity alone does not predict the deficit. A rarer, state-dependent legal class (0.54%) is under-reproduced nearly as much ($U = 0.52$), so the control separates neither state-dependence nor output length from rarity. A linear probe on the residual stream at the last prompt token separates refused from legal actions at AUC 0.81 / 0.84 / 0.94 at 1.7B / 4B / 8B and holds within ± 0.05 of 0.805 at every refusal share of the ladder: the share changes what the model says while leaving this probe unchanged. On 43 byte-identical action pairs that differ only in the state, its pair accuracy is 0.74 / 0.86 / 0.63 (Table D.9). Read like for like, at twenty samples per row, the emission separates refused from legal rows about as well as the probe does (Appendix D.4; Table D.11 is a post hoc exception conditioned on the emission failing), so we find no evidence of knowledge in the state that the emission withholds.

4.3 A second corpus: the cue can be a missing read

The τ-bench retail simulator of §5.3, a Qwen3-1.7B LoRA trained like ours on 4,773 rows from scripted agents with 2.0% refused rows, never reports a refused write ($W = 1.000$ on 200 held-out refused writes), and raising the refusal share to 3–20% leaves W at 1.000 at every rung (Figure 1c).

What limits this simulator is the corpus's construction. We asked the simulator to predict the same 200 write calls behind two prefixes: one in which the agent has read the order the write targets, and one in which it has not (Table 3, left). It refuses *iff nothing was read*, legal and refused alike: its refused rows come only from a scripted agent that writes without reading, so refusal and no-read are perfectly confounded in training and the model learned the cheaper rule; raising the share raises the frequency of the confounded pattern.

Rebuilt from a careless agent that reads the order first and still issues the illegal write, so that the order's status is the only thing separating a refused write from a legal one, the simulator reports the refusals all but completely ($W = 0.000$–0.035) at every share from 2% to 18% (Table 3, right). Below 2% the careless ladder is graded, W falling from 1.000 at 0.2% to 0.000 at 1.5% (one retrained run per rung; Figure 1c).

In both corpora the simulator refuses on whatever feature predicted refusal in its training corpus: the missing read in τ-bench, which no share touches; the share in C10v2, on a cue we have not isolated. The careless corpus reports its refusals from a 1.5% share, the share at which C10v2 whitewashes 0.57, so the share alone does not govern. In C10v2 the cue is not detectably the read: on 21 refused and 37 legal rows asked behind a prefix that reads the decisive fact and one that does not, with a retrained checkpoint that misses its admission tolerance, the refusal rate moves by at most 0.032 and every interval contains zero, an absence of evidence at that size (Appendix D.6).

5 What RL learns from a whitewashing simulator

This section does not measure what the whitewash of §3 costs a policy (no simulator with intermediate whitewash was used as an environment); it asks three narrower questions: whether the simulator's

own reward misorders invalid actions beyond its noise (§5.1); whether a policy follows that reward where invalidity is readable from text (§5.2); and, in a live executor whose simulator never refuses, so that never refusing is the task reward's optimum, how fast GRPO finds it and what restores the refusals (§5.3–§5.4).

5.1 THE SIMULATOR AS A REWARD, AGAINST ITS OWN NOISE

For each battery state we pair a legal and an executor-invalid action and ask which one the simulator's own reward pays more (Bhamidipaty et al., 2026); the reward is the world model's own one-word success verdict on its predicted observation, the pathway a same-model reward design gets for free (Chen et al., 2026). The 8B judge pays the invalid action more on 17–42% of the pairs across the four classes (Table E.1), but on the candidate-set items that misordering sits inside the judge's own ordering noise (unmet-precondition $+0.07$ [-0.04, $+0.19$] above its floor; Table E.3), as does two outside judges' (Appendix E.1): on the battery the self-reward is unbiased within its noise, and §5.2 finds it a slower teacher.

5.2 A POLICY FOLLOWS IT WHERE INVALIDITY IS READABLE FROM THE TEXT

We train a Qwen3-1.7B policy with GRPO to pick one of four candidates per state (two legal, two executor-invalid of one class), so the held-out invalid-choice rate M_1 is exact (67 held-out items in 40 states, 40 of them unavailable-tool; 100 steps; 3 seeds unless stated). An oracle paid by executor legality runs beside every reward as the control.

From the policy's own prior (invalid rate 0.14), the judge reward moves the policy the oracle's way at half its speed: the invalid rate falls by 0.058 against the oracle's 0.121 over 100 steps (ten seeds; difference $+0.063$ [$+0.043$, $+0.084$]); continued to step 300 it covers 0.87 of the oracle's drift, so the gap is one of speed. From a constructed uniform start it leaves 0.25 of held-out choices invalid at 1.7B against the oracle's 0.01 (ten seeds; $+0.24$ [$+0.09$, $+0.39$]; Figure 3a).

A 1.7B classifier separates these candidates at AUC 1.000, so this shows the reward *can* be cashed in where invalidity is decidable from text, not what it costs where it is not. Attempts to build a set no classifier separates failed: from the natural rows (AUC 0.966; the legal member comes from a different state, so a classifier reads coherence rather than legality) and from τ-bench states where only the order's status decides (AUC 1.000; Appendix E.3).

The repair moves the reward only where it moves the wording. On 94 natural-row pairs, a policy trained on the repaired 1.7B's reward ends at M_1 0.394 against 0.713 on the unrepaired 1.7B's (ten seeds a side; -0.320 [-0.543, -0.097]). On the battery, whose tools the bank excludes, the repaired reward leaves the policy within ± 0.03 of the unrepaired one (-0.002 [-0.029, $+0.026$]).

5.3 RL IN A LIVE EXECUTOR

τ-bench retail (Yao et al., 2024) is deterministic: 16 tools and six intents, each requiring an order status. A goal is a customer request on one order, feasible when the status meets the precondition and infeasible otherwise, 1:1; a feasible goal succeeds iff the final database matches the executor's intended write, an infeasible one iff the database is unchanged and the agent's final message says CANNOT. We train a Qwen3-1.7B policy with GRPO in three environments: the real executor; the never-refusing simulator of §4.3 ($W = 1.000$); and that simulator with the executor's refusals passed through as the observation. The reward is the goal's own task, computed from the episode's observations and final message and nothing else (adopted after the registered rule reward failed its own gate; Table E.13): an infeasible goal scores only if some observation was refusal-shaped and the final message says CANNOT; tool identity only, arguments unchecked (Appendix E.4); so inside a simulator that never refuses, never refusing is the optimum.

At step 60 on 96 held-out goals (48 feasible / 48 infeasible; three seeds per cell), real-executor task success S_{real} is 1.000 for the policy trained in the executor, 0.500 (feasible 1.000 / infeasible 0.000) for the one trained in the never-refusing simulator, and 1.000 again with the executor's refusals passed through (Figure 3b–c, totals; the halves in Table E.15). The simulator-trained policy attempts the write on every infeasible goal, is refused live, and reports none of them; passing the executor's refusals through restores the behaviour on every seed under this reward (under the earlier rule reward

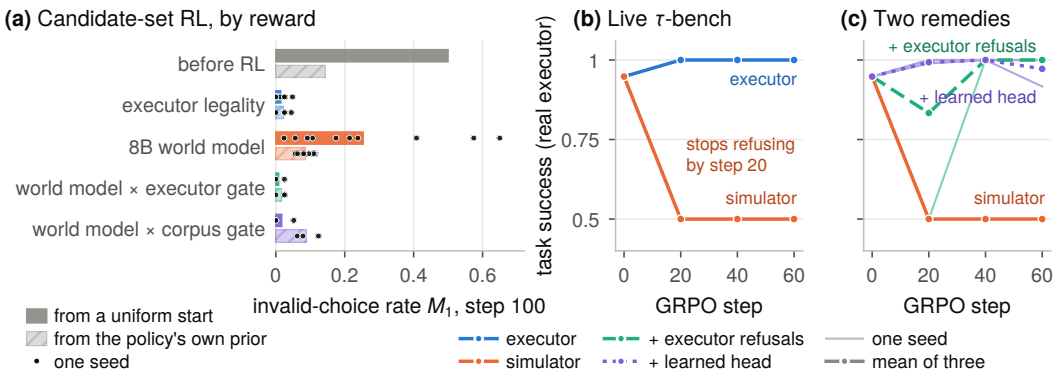


Figure 3: **(a)** Held-out invalid-choice rate M_1 at step 100 for a 1.7B policy in the candidate-set RL, by reward and by where it started; lower is better. Bars are the mean over seeds (ten for the first two rewards, three for the gated rewards of §5.4), each dot one seed (one judge seed collapses onto a fixed candidate); the top row is the policy before RL. **(b, c)** Real-executor task success on the 96 held-out τ-bench goals along training, by the environment the policy trained in: one thin line per seed, the marked line their mean at each evaluated step. All four arms start from the same scripted adapter at 0.948; (b) the executor against the never-refusing simulator, (c) the simulator with each of the two remedies (§5.3, §5.4).

it collapsed to refusing everything; Table E.14). The 0.500 is entailed by the reward once nothing is ever refused, so it is not an effect size; the measured part is the loss from 0.948 to 0.500, on one simulator and one goal set, and the only non-degenerate point we have is the careless simulator of §4.3 at $W = 0.035$, whose policy loses the refusal behaviour on one seed of three (real-executor task success 0.833 against 0.500; Table E.16; under an earlier reward there was no loss to measure, Appendix E.4).

5.4 An admissibility model in the training loop

In the candidate-set RL, V_{state} enters the reward as a multiplicative gate $r \cdot \hat{P}(V{=}1 \mid s, z)$ with the observation untouched; with the executor's own verdict the gated reward ends level with the executor-reward oracle on the candidate sets of §5.2 (Figure 3a).

A learned gate supplies it where invalidity is decidable from text: a classifier trained on the battery's candidates (held-out AUC 1.000) lands the policy beside the oracle from both starts, one trained on the corpus's own refusal rows (AUC 0.766) from the uniform start only (Appendix E.5).

In the live executor the classifier's verdict replaces the *observation*, as the executor's refusals did in §5.3, and the task reward is read from it. A 1.7B classifier trained only on the τ-bench simulator's corpus (its 96 refused rows plus tool-matched legal rows; AUC 1.000 on that corpus, 0.922 on the live goals) rescues nothing at its frozen threshold of 0.5, because its scores saturate ($+0.000$). Re-thresholded after that failure (the re-threshold registered before it ran) at the training goals' executor refusal rate, a base rate into which no held-out goal or arm result enters, it keeps task success at 0.972 [0.951, 0.990] against the simulator-trained 0.500, recovering 0.94 [0.74, 1.17] of the gap to the executor-trained policy without a per-call executor query (Figure 3c). The remedy is not executor-free: the base rate, like the Platt calibration that lands at an adjacent threshold (Appendix E.5), needs the executor's verdicts, which the corpus does not supply.

6 Related work

Learned environments. DreamGym, the Patronus corpus, a masked-diffusion world model, and Qwen-AgentWorld (Chen et al., 2026; Patronus AI, 2026; Deshpande, 2026; Zuo et al., 2026) train next-observation simulators and optimise policies in them, scoring fidelity on legal actions; surveys and benchmarks of text-world simulators (Li et al., 2026a;b; Liu et al., 2026b) catalogue hallucinated entities and state drift. AgentWorldBench's "Error Handling" axis scores whether a deterministic-

error turn is answered with an error signal: the phenomenon is not new with us. That axis credits any error signal on turns selected by outcome and sits downstream of the error-turn filter; ours reads for success narration, excludes program exceptions, and adds a battery that selects by action validity (Appendix C.4). Programmatic environments (Song et al., 2026) and memory-augmented simulators (Wang et al., 2026a) attack fidelity structurally. Relative to that axis, we add where the deficit sits (Table C.1), what moving it costs, and what restores a policy once the omission is complete.

Emulators and the policy side. Prompted emulators (Ruan et al., 2024) simulate tool outputs, including the exceptions a tool's specification names, from a description rather than from an executor; we measure fine-tuned simulators against the executor's own verdicts, and our prompted arm carries the flagship's format with no exception specification, so it does not test that recipe. Advani (2026) measures the policy's false success claims, of which whitewash is the environment-side counterpart; a bare LLM self-check hallucinates validity for impossible actions (Liu et al., 2026a). Oefinger et al. (2026) argue a world model used as an oracle must be accredited first, naming the action-coverage gap; we measure it with an executor in the loop.

Model error and the optimiser. Offline model-based RL penalises leaving the data's support (Sutton, 1991; Janner et al., 2019; Yu et al., 2020; 2021); an optimiser on a rich policy class finds a learned model's errors (Bhamidipaty et al., 2026); reward hacking prices a proxy reward wrong where it is optimised (Skalse et al., 2022; Gao et al., 2023). Action masking and shielding (Zahavy et al., 2018; Huang & Ontañón, 2022; Alshiekh et al., 2018) act on the policy; coded state (Lu et al., 2026; Wang et al., 2026b) supplies admissibility by construction.

7 Limitations

The cause of the deficit remains open. Both corpora show the simulator refusing on the cue that predicted refusal in training; we have not isolated C10v2's cue, and the reading that fails to find the read there is 21 refused rows on a checkpoint that misses its admission tolerance (Appendix D.6). On the 42% of natural rows the decidability read calls undecidable, W measures what the simulator reports, not what it could infer; that read is one LLM reader, validated by one author on 100 rows, and both saw the recorded refusal, so it grades explicability after the fact rather than prediction (Appendix C.6). The four class labels on the natural rows are the authors' rules over template groups, not independently validated. The instrument reads reported refusals; a state changed wrongly with no error is outside it. Every interior ladder rung, and the upsampled-own-rows cell of the 2×2, is one training run against a 1.7B seed SD of 0.092; the ladder's endpoints, the two banks at 1.5% and the scale trend are replicated at three seeds, and two of the unreplicated contrasts fall inside that SD; the high-share arms also saw 23% more training rows, a budget we did not hold fixed (Appendix D.2). The RL results are existence proofs: a 1.7B classifier separates all three candidate sets, so §5.2 measures where the reward can be cashed in, and §5.3 measures the loss at $W = 1.000$ only; no policy was trained in a simulator that whitewashes part of the refusals (arms in the retrained $W = 0.55$ and 0.17 simulators were registered and voided when the retrained warm start failed its registered acceptance rule; Appendix E.4), and other goal sets and rewards are untested. Our own models are 1.7B–8B LoRA adapters on a 30k-row slice; the released model, whose pipeline is described as removing error turns, whitewashes least of the models we read, so scale and data volume, which we do not vary, may matter more than the share; whether the deficit closes with scale our data bound but do not answer. The human read covers 220 rows and 46 flagship items, and 77 registered readings carry no family-wise correction.

8 Conclusion

Three things to take away. *Measure it:* fidelity on legal actions does not test admissibility; score the simulator on actions the executor refused and read the samples for success narration—every model we audit whitewashes, the pretrained checkpoints already do. *Adding refusals buys emission, not admissibility:* upsampling the corpus's own refusals lowers whitewash at a false-refusal price (§4), leaves a probe of the state unchanged and reaches almost none of the planted state violations; a two-prefix test tells a rule from a cue. *Do not train a policy inside a simulator that never refuses:* the reward makes never refusing optimal and GRPO finds it in twenty steps; only the executor's verdict, per call or through a learned head thresholded on it, restores the behaviour.

AI USE STATEMENT

(Required by ICLR 2027; does not count toward the page limit.) Generative AI tools (large language models used as coding and writing assistants) were used for tasks with required disclosure: proposing and refining hypotheses and experimental designs (the registrations in Appendix G were drafted with LLM assistance under the authors' protocol: thresholds frozen before data, registration texts dated before their data, with the exceptions disclosed in Appendix G.1), implementing the methods (training, probing, battery-construction and analysis code), cleaning and reformatting data, supporting qualitative analysis (the blind agent reads of §3.1 and Appendix C.4–C.6 were produced by LLM readers under the registered protocols; they are labelled as agent reads wherever quoted and are not human validity—both registered human reads, of the 220-row whitewash sheet and of the flagship's own-benchmark sheet (Appendix B.5), were read by an author and by a second, paper-naive reader, and are reported as human reads; the 100-row decidability read (Appendix C.6) was made by an author alone and serves only to validate an LLM reader's labels; that one reader is an author is stated as a limitation (§7, Appendix B.5)), and interpreting results. We did not use generative AI tools to formulate or prove mathematical claims (the paper makes none), and synthetic-data generation is not applicable beyond the script-constructed battery variants of §2, whose executor-validation is mechanical. Additionally, generative AI tools were used for tasks with recommended disclosure: drafting and editing the text, making figures, editing code, and literature analysis. We have reviewed all AI-assisted work: every number in the paper is produced by a script in the repository from a stored readout and was re-derived in three internal review passes; quotations from cited work were checked against the sources (§6); LLM-written code was tested against executor ground truth (Appendix A.3, B.1). The authors set the research question, reviewed the registrations, analyses and text, and take responsibility for the final content, including text, claims and artifacts produced with the aid of generative AI. The flagship's own-benchmark agent read agrees with two human readers on a one-sample-per-row census (46 items, agreement 0.98 and 0.98; Appendix B.5), the second of whom had never seen the paper.

REPRODUCIBILITY STATEMENT

Training, the C10v2 world-model arms. LoRA $r = 16$, $\alpha = 32$, dropout 0.05 on all attention and MLP projections; one epoch, batch $1\times$ grad-accum 16×2 devices, lr 1e-4 cosine with 3% warmup, bf16, sequences capped at 16,384 tokens (`method_train.py`; its parent `p5_train_8b_wm.py`, the same recipe, trained the first seed of the three unaugmented arms). The Y-harvested dose 2 trains two epochs (Appendix B.2). The τ-bench simulators, including every careless-ladder rung, use the same recipe on one device at 4,096 tokens for two epochs (`p5_train_8b_wm.py`). The warm start and the two learned gates are $r = 8$, $\alpha = 16$, dropout 0, on q, k, v, o only, three epochs at lr 1e-4, batch 8 (`rle_warmstart.py`; the gates by `rlk_gate.py`; Appendix E.5), and the GRPO policies train that $r = 8$ adapter at lr 2e-4 with KL 0.02 and no ratio clipping (`rle_train.py`). The optimiser is AdamW throughout and no script sets its hyperparameters: the SFT runs take the trainer's defaults (weight decay 0); the warm start, the gates and GRPO take `torch`'s, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, weight decay 0.01, with gradients clipped at norm 1.0.

Sampling, the 399 natural rows. Three emissions per row at each checkpoint's released generation config (temperature 0.6, top-p 0.95, top-k 20), a 1,024-token observation budget in a 40,960-token context (SmolLM2: 8,192; the base arms and the flagship under their own contract at 16,384) (Appendix F). Elsewhere: five samples per state on the battery, five per row on the flagship's own benchmark, twenty for Appendix D.4's like-for-like comparator and none for the probe, which reads a hidden state; the τ-bench simulator is read once per row at a 384-token budget, and the RL policies train at temperature 1.0 and are evaluated greedily with a 96-token action budget (`rle_train.py`).

Every number is produced by a script in the repository from a stored readout (Appendix G.2 lists, with its source file and key, every main-text number that no table carries and that the surrounding text does not itself derive); every registration is a dated file, available on request (Appendix G.1). We release, in an anonymised repository (https://anonymous.4open.science/r/whitewash-audit-6635), the training and probing code, the frozen batteries (the 1,336-variant battery, its four invalidity recipes execution-validated and no item itself executed, with its justifications, the 399-row natural-refusal set, the 60-row flagship set (48 primary rows scored; Appendix C.4)), the readouts behind every table and figure, and the stored emissions the

natural-refusal audit and the flagship read are computed from (one file above the host's size limit is listed there by hash); the read sheets and labels are in the supplementary material. The LoRA adapters are not released: most of the weights that produced these emissions no longer exist. Every printed number is read from a stored readout in the repository, and the natural-refusal and flagship readings can be recomputed from the stored emissions; new counterfactual queries to the original checkpoints cannot be run, only to retrained ones. The exception is a registered retraining: the four sub-2% τ-bench simulators and the warm start were retrained, and §4.3's and §7's readings of that ladder come from the retrained weights, which differ from the originals where both were read (Table A.3).

REFERENCES

Laksh Advani. From confident closing to silent failure: Characterizing false success in LLM agents. In *ICML 2026 Workshop on Foundations of Agentic Systems (FAGEN@ICML 2026)*, 2026. URL https://arxiv.org/abs/2606.09863.

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2669–2678, 2018. doi: 10.1609/aaai.v32i1.11797. URL https://arxiv.org/abs/1708.08611.

Logan Mondal Bhamidipaty, Esmeralda S. Whitammer, David Abel, Mykel J. Kochenderfer, and Subramanian Ramamoorthy. Imperfect world models are exploitable. *arXiv preprint arXiv:2605.15960*, 2026. URL https://arxiv.org/abs/2605.15960.

Zhaorun Chen, Zhuokai Zhao, Kai Zhang, Bo Liu, Qi Qi, Yifan Wu, Tarun Kalluri, Sara Cao, Yuanhao Xiong, Haibo Tong, Huaxiu Yao, Hengduo Li, Jiacheng Zhu, Xian Li, Dawn Song, Bo Li, Jason Weston, and Dat Huynh. Scaling agent learning via experience synthesis. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2511.03773. DreamGym.

Darshan Deshpande. Masked diffusion language models are strong and steerable text-based world models for agentic RL. *arXiv preprint arXiv:2607.16204*, 2026. URL https://arxiv.org/abs/2607.16204. Dataset: https://huggingface.co/datasets/PatronusAI/world_model_corpus; code: https://github.com/patronus-ai/mdlm_world_modeling.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 10835–10866, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Shengyi Huang and Santiago Ontañón. A closer look at invalid action masking in policy gradient algorithms. In *Proceedings of the International FLAIRS Conference*, volume 35, 2022. URL https://arxiv.org/abs/2006.14171.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://arxiv.org/abs/1906.08253.

Yixia Li, Hongru Wang, Peng Lai, Zhiwen Ruan, He Zhu, Youxin Zhu, Ganlong Zhao, Minda Hu, Yun Chen, Sibei Yang, Peng Li, Jeff Z. Pan, Jia Pan, Guanhua Chen, Yang Liu, and Guanbin Li. Bridging the agent-world gap: Text world models for LLM-based agents. *arXiv preprint arXiv:2606.09032*, 2026a. URL https://arxiv.org/abs/2606.09032.

Yixia Li, Hongru Wang, Jiahao Qiu, Zhenfei Yin, Dongdong Zhang, Cheng Qian, Zeping Li, Xiaoteng Ma, Guanhua Chen, and Heng Ji. From word to world: Can large language models be implicit text-based world models? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8084–8111, 2026b. doi: 10.18653/v1/2026.acl-long.366. URL https://aclanthology.org/2026.acl-long.366/.

Pengxiang Liu, Tao Ren, Wei Xiong, Tingrui Yang, Junjie Wang, and Jun Hu. OntoGuard: Enforcing action admissibility for LLM agents in complex interactive environments. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20937–20952, San

Diego, California, United States, 2026a. doi: 10.18653/v1/2026.findings-acl.1051. URL https://aclanthology.org/2026.findings-acl.1051/.

Yi Liu, TingFeng Hui, Wei Zhang, Li Sun, Ningxin Su, Jian Wang, and Sen Su. EnvSimBench: A benchmark for evaluating and improving LLM-based environment simulation. *arXiv preprint arXiv:2605.07247*, 2026b. URL https://arxiv.org/abs/2605.07247.

Xiaoxin Lu, Ranran Haoran Zhang, and Rui Zhang. SIMMER: Benchmarking latent failures in LLM executable planning with a world model. In *Conference on Language Modeling (COLM)*, 2026. URL https://arxiv.org/abs/2606.14574.

Christian Oefinger, Finn Rasmus Schäfer, Korbinian Moller, Mattia Piccinini, and Johannes Betz. Validate the dream before you trust its verdict: Admissibility for world-model simulators. In *RSS 2026 Workshop on Robot World Models*, 2026. URL https://arxiv.org/abs/2607.07196.

Patronus AI. PatronusAI/world_model_corpus. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/PatronusAI/world_model_corpus. 239,403 train / 19,454 val / 20,847 test items over nine environments; MIT license.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.15817. Spotlight.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2209.13085.

Xiaoshuai Song, Haofei Chang, Guanting Dong, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. EnvScaler: Scaling tool-interactive environments for LLM agent via programmatic synthesis. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL https://aclanthology.org/2026.findings-acl.407/.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991. doi: 10.1145/122344.122377.

Yujun Wang, Tao Zhang, Jinhe Bi, Aniri, Wenxuan Ye, Boliang Liu, Sikuan Yan, Shuning Wang, Xuebing Zhou, Sören Pirk, Hinrich Schütze, and Yunpu Ma. MemWM: Memory-augmented text-based world model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2026a. URL https://arxiv.org/abs/2608.07107.

Zhaoyang Wang, Canwen Xu, Boyi Liu, Yite Wang, Siwei Han, Zhewei Yao, Huaxiu Yao, and Yuxiong He. Agent world model: Infinity synthetic environments for agentic reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026b. URL https://arxiv.org/abs/2602.10090.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020. URL https://arxiv.org/abs/2005.13239.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. COMBO: Conservative offline model-based policy optimization. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021. URL https://arxiv.org/abs/2102.08363.

Tom Zahavy, Matan Haroush, Nadav Merlis, Daniel J. Mankowitz, and Shie Mannor. Learn what not to learn: Action elimination with deep reinforcement learning. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018. URL https://arxiv.org/abs/1809.02121.

Yuxin Zuo, Zikai Xiao, Li Sheng, Fei Huang, Jianhong Tu, Yuxuan Liu, Tianyi Tang, Xiaomeng Hu, Yang Su, Qingfeng Lan, et al. Qwen-AgentWorld: Language world models for general agents. *arXiv preprint arXiv:2606.24597*, 2026. URL https://arxiv.org/abs/2606.24597. Models Qwen/Qwen-AgentWorld-35B-A3B and 397B-A17B, dataset Qwen/AgentWorldBench, released 2026-06-24; https://github.com/QwenLM/Qwen-AgentWorld.

A DATA

This appendix expands the main text in the order a reader is likely to check it: the data (A), the instruments and their validation (B), the full results behind §3, §4 and §5 (C–E), training and sampling details (F), and the pre-registration ledger with the provenance of every number (G). Every table is produced by a script in the repository from a stored readout; Appendix G.2 indexes the ones the main text quotes without a table. The read sheets and labels behind Appendix B.5 and C.6 are in the supplementary material (G.3).

A.1 THE C10V2 CORPUS AND THE TRAINING SLICE

Supports §2 (*Corpus, models, training*) and the 1.5% refusal share of §1 and §4.1. C10v2 is the Patronus world-model corpus: 239,403 train / 19,454 eval / 20,847 test items over nine environments, each item one recorded (trajectory prefix, action, observation) row. The head rule of Appendix B.1 scanned every row whose last message is a recorded observation of a tool action; rows whose "action" is a tool or user message are a corpus quirk and were skipped. A head-rule hit whose first line is a program, HTTP or SQL exception rather than the executor refusing is counted separately and excluded.

Table A.1: The census of verified executor refusals in C10v2, by split. The train row is reproduced by `scripts/c10v2_scan.py` (`outputs/c10v2/scan_train.log`); the eval and test rows are the same scan on those splits.

split	rows	head fires	of which a program/HTTP/SQL exception (`other`, excluded)	verified executor refusals	share of rows
train	239,403	3,664	413	3,251	1.36%
eval	19,454	414	55	359	1.85%
test	20,847	392	51	341	1.64%

Train-split refusals are 2,863 API-tool and 388 SWE rows over 216 templates; the ten most frequent templates cover 60% of them and the five most frequent 41%. Every registered arm saw the same 30,000-row slice of the train split (the training-seed replicates of Table D.2 re-draw the subsample with the seed; Appendix F) (seed 20260826, `idx[500:30500]` after the seeded shuffle; `scripts/c10v2_exposure_split.py`), of which 444 rows begin with an executor error (1.5%); the 16,384-token length filter drops about 460 further rows at training time.

A.2 THE NATURAL REFUSAL SET AND ITS LEGAL COMPANIONS

Supports §2 (*The natural refusal set*) and every number of §3.1. From the eval and test splits — out of sample for every model we train — we froze, before any model ran, 399 refused rows after a cap of ten per refusal-template group (seed `c10v2|20260902`; sha256 `1ea58b0b...c405`): nonexistent-entity 178 (SWE 2, API 176), unmet-precondition 92 (SWE 10, API 82), invalid-args 68 (SWE 20, API 48), unavailable-tool 61 (API), over 161 template groups. The API tools span vehicle, file-system, trading, travel, telecom, retail and MCP surfaces; the SWE rows are OpenHands editor and shell actions. Class rules over the (class, tool, template) groups were hand-read before freezing. Selecting rows with the same rule that reads the emissions biases the set *toward* recovery — a model that reproduces the corpus's own error line is credited — so the selection works against the whitewash claim, not for it.

The legal companion is 185 tool-matched eval/test rows whose recorded observation has no error head and no traceback (53 template groups; sha256 `2f7cf280...9e28`). No state carries both a refused and a legal row (0 of 399), which is why the RL candidate sets of Appendix E.3 are tool-matched

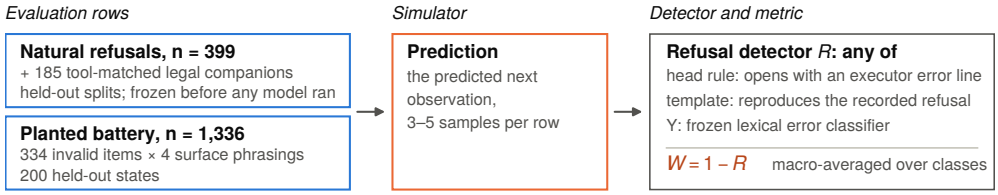


Figure A.1: What is measured, on what, and with which instrument. The natural rows are the executor's own recorded refusals; the battery's variants are built by four recipes validated on a live executor and are not themselves executed (A.3); both are out of sample for every model we train. The detector is the union of three rules and is deliberately permissive, so that the simulator gets credit for any refusal signal; whitewash is $W = 1 - R$, macro-averaged over the four refusal classes with a template-clustered bootstrap (B = 2,000). The companions drop the template rule, which on a legal row detects fidelity rather than refusal, and the battery uses the head rule alone, since Y fires on file listings.

rather than state-matched. **Where the observations come from.** Every one of the 399 recorded refusals is attributable to its source: 32 are live executions (the SWE rows), 42 are recorded retrievals from the corpus's API tools, and 325 come from its deterministic mock tools; none is a language model's output (the per-row provenance table is in the supplementary material).

The cap, priced. Weighting each of the 161 template groups by its count in the 700-refusal census of the held-out splits, instead of capping at ten, gives a deployment-frequency reading of the same stored emissions (`scripts/returns_freqw.py`; template-clustered bootstrap, B = 2,000). It lowers W for every arm: Qwen3-1.7B 0.525 → 0.492 [0.404, 0.589], 4B 0.456 → 0.419 [0.351, 0.504], 8B 0.263 → 0.222 [0.182, 0.283], SmolLM2 0.676 → 0.646, the released 35B 0.253 → 0.201, the prompted 30B 0.497 → 0.373, the 724-row repair 0.181 → 0.181 and the upsampled-own-rows arm 0.208 → 0.181. The scale ordering, the repair's level and every conclusion of §3–§4 are unchanged; the frequent templates are the ones the models reproduce best, so the capped set is the harder reading.

A.3 The planted battery

Supports §2 (*The planted battery*) and §3.2. The 200 states are rows of the corpus's test split, chosen with a fixed seed after stratifying by the parsed tool's domain (at least 50 per domain; `scripts/p4_prep.py`); every trained model's slice is drawn from the train split, so no battery state was seen in training. On these 200 states (100 SWE, 100 API-tool) the battery plants three tiers of variants of the recorded action z_0. Tiers 1–2 are E-equivalent re-serialisations and register-controlled decorations (guard: the perturbed action parses to the same canonical call as z_0); Tier 2.5 is the executor-invalid tier the paper reads — parseable, well-formed, refused — 334 items in four classes over four surface registers, 1,336 variants, each carrying the recipe's justification of why the executor would refuse it, graded airtight / strong / presumptive and checked for presence (0 guard failures). Unavailable-tool and invalid-args violate conjuncts a static schema check can shield cheaply (the static classes); nonexistent-entity and unmet-precondition violate the state conjunct a deployed simulator must own. Three classes exist on SWE states only (44–45 items each); unavailable-tool has 199. One item's state exceeds every context window, so the Qwen arms read 333 of 334 items.

The four recipes (`scripts/t25_build.py`), each applied to the parsed canonical action and each checked against the whole state before the item is kept. *Nonexistent-entity*: one identity-bearing string argument whose value occurs verbatim in the state text is replaced by a planted token `nx_<8 hex>` (appended to the path when the key is a path), asserted absent from the system prompt, the prefix and the recorded observation; graded presumptive, because an API may answer an unknown id with a well-formed not-found payload rather than an error; the items kept all fall on SWE states (editor views of a planted path), where the editor answers with an error. *Unavailable-tool*: the tool name is replaced by the most frequent tool in the battery's census that is not this state's tool and does not occur in the state text, arguments unchanged; dispatch on a name the state's tool surface does not offer fails structurally (airtight). *Invalid-args*: for the editor, `view` with `view_range = [10000000, 1]`; otherwise one count-like integer argument set to -2^{31}; range-invalid by the parameter's own type (strong). *Unmet-precondition*: `str_replace` on the canonical path with `old_str` a planted

token `ZQXNEVER<8 hex>` asserted absent from the state, so the editor's not-found path is the only reachable one (airtight).

Ground truth by execution. The same four recipes, instantiated on the tools of a live deterministic environment (τ-bench retail) and run there, error on 240 of 240 executions and cover 37 of 37 of the executor's rejection paths. This validates the recipes, not the items: no battery item is itself executed, because its state comes from a trajectory that cannot be re-run.

The admission gate. Each model must reach canonical-direction accuracy ≥ 0.80 on the 199 states' recorded legal actions (five samples per state; the predicted observation carries the recorded observation's outcome label under the frozen classifier Y) before the battery is read. Table A.2 re-scores the same emissions under the refusal instruments of Appendix B.1: the two refusals the gate's states hold and the 197 accepted actions.

Table A.2: The admission gate re-scored under the head $\vee$ template rule (`scripts/gate0_rescore.py`). Y-direction is the registered gate quantity, reproduced to four decimals; the last two columns are the majority verdict and the sample share on the two refusals and on the 197 accepted actions.

model / run	states	Y-direction (registered)	head $\vee$ tmpl agreement	recovery on the 2 refusals (majority; samples)	phantom refusal on 197 accepted (majority; samples)
Qwen3-8B, G0-B run	199	0.9146	0.995	2/2; 0.90	0.005; 0.008
Qwen3-8B, T25 run (§2 verdict)	199	0.9095	0.995	2/2; 0.90	0.005; 0.008
Qwen3-4B	199	0.8995	0.995	1/2; 0.50	0.000; 0.001
Qwen3-4B, cross-model run	199	0.9045	0.995	1/2; 0.60	0.000; 0.003
Qwen3-1.7B	199	0.8693	0.995	1/2; 0.50	0.000; 0.003
Qwen3-1.7B, cross-model run	199	0.8693	0.995	1/2; 0.50	0.000; 0.003
SmolLM2-1.7B (E3 run; 144 states in context)	144	0.9028	0.993	1/2; 0.50	0.000; 0.004
AgentWorld-35B-A3B, own simulator prompt	199	0.7739	0.990	0/2; 0.00	0.000; 0.002
AgentWorld-35B-A3B, + role-lock repair	199	0.7739	0.990	0/2; 0.00	0.000; 0.001

The refusal every SFT arm recovers is the one whose template is fixed for the tool (`{"error": 404}`, 5/5 samples on every arm); the one only the 8B recovers is the state-dependent one, where the 1.7B, 4B and SmolLM2 arms print the file listing a valid `view_range` would have produced. The released flagship scores 0.774 under both of our prompt contracts and is therefore read only on the natural rows, in its own contract (Appendix C.4).

A.4 The two τ-bench corpora

Supports §4.3 and §5.3. τ-bench retail is deterministic: 1,000 orders, 16 tools, six intents (cancel and the three modify intents require a *pending* order; return and exchange a *delivered* one). The base simulator corpus holds 4,773 rows from competent, thorough and sloppy scripted agents acting in the real executor, with refused rows fixed at 2% $\pm$ 0.5% (landed 2.01%). Its refused rows come only from the sloppy agent, which writes without first reading the order, so in this corpus every refusal sits behind a prefix with no read. The *scripted* ladder adds sloppy episodes on infeasible goals until refused rows are 3–20% of the corpus. The *careless* ladder rebuilds the corpus from an agent that reads the order first and still issues the illegal write, so every refused row sits behind the same read a legal write does and the order's status in the read observation is the only thing that separates them; its rungs run from 0.2% to 18% (the 20% target is capped by the 1,343-goal infeasible pool). Each simulator is a Qwen3-1.7B LoRA trained by the recipe of Appendix F, read once per row on 200 held-out legal and 200 held-out refused write calls; the legal-write pass rate is the validity gate ($\geq$ 0.80).

Table A.3: The two τ-bench ladders: whitewash W on the 200 held-out refused writes and the legal-write pass rate, one simulator per rung. The four sub-2% careless rungs are retrained checkpoints (Appendix F); the original checkpoints read 1.000 / 0.655 / 0.130 / — . Figure 1c plots the W columns.

refused share of corpus rows	scripted corpus: W	scripted: legal pass	careless corpus: W	careless: legal pass
0.21%	—	—	1.000 (retrained)	1.000
0.51%	—	—	0.550 (retrained)	1.000
1.0%	—	—	0.170 (retrained)	1.000
1.5%	—	—	0.000 (retrained)	1.000
2% (the base corpus)	1.000	1.000	0.035	1.000
3%	1.000	1.000	—	—
5%	1.000	1.000	0.000	1.000
8%	1.000	1.000	0.000	1.000
12%	1.000	1.000	0.000	1.000
18–20%	1.000 (20%)	1.000	0.000 (18%)	1.000

B Instruments and their validation

B.1 The rules, verbatim

Supports §2 (*Instruments*). All rules read the emission after `STRIP = ^\s*<tool_response>\s*(?:OBSERVATION:\s*)?` is removed.

Head rule R_{head}: `HEAD` matches the first 80 characters of the body — `^(?:Error:|ERROR:|\{\s*"error"|bash: |/bin/(?:ba)?sh: |\S+: (?:command not found|No such file)|Tool \S+ not found)` (case-insensitive). For shell tools (`execute_bash`, `bash`, `terminal-run_command`, `run_command`, `shell`, `run_shell`) a program's own `Error:` print is not the executor refusing, so the head must additionally match `SHELL_HEAD = ^(?:bash: |/bin/(?:ba)?sh: |\S+: (?:command not found|No such file)|Error: Command execution failed)`. On legal actions the rule fires at 0.008–0.017 across the four SFT arms (Table 2's legal-baseline entries); on the 399 verified natural refusals it fires on 399/399 recorded observations by construction of the row set.

Template rule R_{tmpl}: `tmpl_key(obs)` unwraps a BFCL-style `{"error": "..."}` wrapper, drops a lone `ERROR:` first line, truncates an unclosed backtick span, replaces quoted spans, numbers and paths by ·, lower-cases and keeps the first 40 characters of the first non-empty line. The rule fires when the row's recorded-observation key (≥ 12 characters) occurs in the normalised first 3,000 characters of the emission. It credits a refusal that is the corpus's own template even when the head differs. On a *legal* row the same term detects that the model reproduced the normal recorded observation — fidelity, not refusal — which is why false refusals on the legal companions are read with head $\vee$ Y (Table G.1).

Y is a frozen lexical outcome classifier (label `error`): a first-match-wins list of case-insensitive regular expressions, nothing learned (`agentrl_wm/outcome/y_classifier.py`, bank `wmc`). Its `error` pattern matches any of `error during execution`, `"error":`, `error:`, `traceback (most recent`, `unexpected keyword argument`, `exception`, `not found`, `invalid`, `permission denied`, `failed`, `failure`, `assertionerror` as whole words; a `success` pattern (`"status": "success|ok|completed"`, `successfully`, `"success": true`, `all tests passed`, `passed ... 0 failed`) and a `neutral` fallback follow. On the natural rows $R = R_{\text{head}} \vee R_{\text{tmpl}} \vee R_Y$ is deliberately permissive — the world model gets credit for any refusal signal — and $W = 1 - R$. Y keeps three labelled roles, each with its bias direction stated: as a disjunct of R, where a Y false positive can only credit the model (lowering W, against our claim); as the lexical checker channel of §5.1, where its 0.06–0.12 false-positive rate on legal rows is the baseline that produces the ties; and as the label of every Y-positivity reading (the Y column of Table B.3). On the battery, where Y fires on file listings (B.2), the head rule is read alone.

Class-consistency rule r_{cc} (analysis instrument, battery only): the head fires *and* the first 600 characters name the element the mutation planted — the renamed tool; `10000000`, or `2147483648` together with the parameter it was assigned to; the `nx_[0-9a-f]{8}` entity id; the `ZQXNEVER[0-9a-f]{8}` old-string — or use that class's executor vocabulary. A refusal that fires

the head but not r_{cc} is a refusal for the wrong reason. **`cat -n`** is the OpenHands editor's listing preamble (`Here's the result of running` cat -n); it is reported as a descriptive column wherever Y is, because it is what Y was mostly counting.

B.2 What Y counts, and the error bank it harvested

Supports §4.1 (*The repair*, the failed Y-harvested doses). The first repair paired each synthesised executor-invalid action with a target drawn from an "error bank": every recorded training observation for the same tool that Y labelled `error`. Applying the head rule to the two banks shows what they hold.

Table B.1: What the Y-harvested error banks contain, by the head rule.

bank	rows	begins with an executor error	`cat -n` file listing	other normal output
dose 1 (`augment.jsonl`)	1,735	304 (17.5%)	1,192 (68.7%)	239 (13.8%)
dose 2 (`augment_dose2.jsonl`)	7,278	1,016 (14.0%)	5,336 (73.3%)	926 (12.7%)

"Other normal output" is directory listings, `The file ... has been edited` success messages, test banners and `submit` checklists; Y labelled every one of them `error` because the body contains `Error`, `Exception` or `invalid` somewhere in source code, a diff or a test name. The registration's own check — "labels are real error observations (frozen Y agrees on 100%)" — was circular with the harvest. The models trained on these banks learned exactly what they were shown:

Table B.2: What the Y-harvested arms emit on the battery's three decorated registers (999 variants × 5 samples = 4,995 emissions per arm).

arm	emissions beginning with an executor error	emissions that are `cat -n` listings
unmodified 1.7B	58 (1.2%)	1,894 (38%)
dose 1	145 (2.9%)	3,908 (78%)
dose 2	183 (3.7%)	3,367 (67%)

Y fires on 22–56% of those listings (by class) against 1.00 on genuine error heads, which is the whole of the preliminary Y-read "repair" signal. The repair that works (dose-v2, §4.1) is built from natural refusals selected by the head rule (`augment_v2.jsonl`, 724 rows, sha256 `c83ad58a...22ba`), class-filtered to unavailable-tool and invalid-args, held-out tools excluded, capped at 60 per template group, and verified before training.

B.3 A wrapper-agnostic detector

Supports §2 (*Instruments*) and §3.1 (the base checkpoints). The head rule was written against the corpus's own error line. A structural, format-agnostic detector — the head rule, or a parsed JSON object carrying an error-family key at any depth (a value in {error, failure, failed, denied, refused}, a true boolean, or a non-zero exit code), or the same wrapper written as a bare `error:` line — was registered before it was scripted and read on the same rows.

On the seven arms trained on this corpus the detector changes macro-W by at most 0.002 and costs at most +0.020 in false refusals: their refusals are in a format the frozen rule already parses. What it recovers is on the checkpoints that never saw the corpus — the base models and, less, the flagship — which report refusals inside their own `{"type": "error", ...}` wrapper. Over the 44 arms whose readouts carry per-row instrument columns, macro-W under R and under Y alone differ by at most 0.007, while the head rule alone reads up to 0.210 higher: every W the paper prints is the conservative reading.

B.4 The 28-row audit of the preliminary instrument

A preliminary version of the natural-row measurement rested on 28 corpus actions selected by a class rule together with Y labelling the recorded observation `error`. Reading the 28 recorded observations one by one found 5 genuine refusals (selection precision 0.18) and the class wrong on

Table B.3: The frozen head rule against the head rule plus the wrapper detector, on the same rows, with the registered union R beside them; Δ is head minus wrapper, so a positive Δ is a fall in whitewash. The last two columns are the wrapper clause's cost on the 185 legal companions and the share of samples on which it fires where the head rule does not. Template-clustered bootstrap, B = 2,000.

arm	macro-W, head only	macro-W, head $\vee$ wrapper	Δ (head $-$ wrapper)	registered R	R's Y term alone	legal FP, head $\rightarrow$ wrapper	wrapper fires beyond head
1.7B	0.616	0.616 [0.538, 0.695]	+0.000	0.525	0.527	0.045 $\rightarrow$ 0.049	0.0%
4B	0.538	0.537 [0.459, 0.619]	+0.000	0.456	0.459	0.022 $\rightarrow$ 0.022	0.1%
8B	0.344	0.343 [0.278, 0.415]	+0.000	0.263	0.265	0.038 $\rightarrow$ 0.038	0.1%
SmolLM2	0.801	0.799 [0.743, 0.853]	+0.002	0.676	0.677	0.043 $\rightarrow$ 0.043	0.2%
dose 1	0.671	0.670 [0.605, 0.735]	+0.001	0.599	0.601	0.041 $\rightarrow$ 0.041	0.2%
dose 2	0.686	0.685 [0.616, 0.756]	+0.001	0.607	0.613	0.036 $\rightarrow$ 0.040	0.2%
dose-v2	0.206	0.206 [0.144, 0.279]	+0.000	0.181	0.181	0.083 $\rightarrow$ 0.083	0.0%
1.7B base	0.927	0.786 [0.720, 0.847]	+0.141	0.760	0.762	0.013 $\rightarrow$ 0.025	14.0%
4B base	0.840	0.627 [0.558, 0.712]	+0.212	0.647	0.650	0.002 $\rightarrow$ 0.022	24.5%
8B base	0.726	0.693 [0.630, 0.783]	+0.032	0.516	0.516	0.002 $\rightarrow$ 0.013	3.5%
the flagship	0.440	0.415 [0.336, 0.502]	+0.024	0.253	0.259	0.000 $\rightarrow$ 0.011	2.6%

4 of the 5; on the 23 non-refusal rows a world model that predicts a normal observation is *correct*, so the preliminary signed statistic mostly measured agreement with the corpus. The preliminary measurement was set aside, the head rule replaced Y as the selecting instrument, and the 399-row set of A.2 was built from the resulting census.

Table B.4: The 28 preliminary rows by class, with the recorded observation and whether it is a genuine executor refusal.

rows	k class	tool	recorded observation	genuine executor refusal?
0	unavailable-tool	`canvas_get_user_profile`	`Tool canvas_get_user_profile not found in agent Assistant`	**yes**, class correct
1–8	unavailable-tool	`submit`	"Thank you for your work on this issue…" checklist + diff	no — the diff text contains `Exception`, `invalid`, `Error`
9, 25	unavailable-tool	`execute_bash`	a script's own runtime output (`... ERROR:`, `error: invalid command 'test'`)	no — the command ran; the program printed
10, 11, 17, 24, 27	unavailable-tool	`str_replace_editor`	`Here's the result of running cat -n ...`	no — a file listing
12–15, 18–20, 23	nonexistent-entity	`str_replace_editor`	`Here's the result of running cat -n ...`	no — a file listing
16, 21, 22, 26	unavailable-tool / nonexistent-entity	`str_replace_editor`	`ERROR: Invalid view_range parameter ...`	**yes**, but the class is invalid-args

B.5 THE HUMAN READS

Supports §3.1 (the human check of the instrument; the released flagship). **The 220-row sheet.** `scripts/c10v2_read_sheet.py build` draws 20 rows per landed arm (5 per class, by hash; 40 for dose-v2), one sample each, shuffles them with a fixed seed and writes the sheet without model identity. Two readers — reader 1, an author; reader 2, who had never seen the paper, the key files or reader 1's labels — labelled every row *refusal reported*, *whitewash* or *other*. The registered rule (registered before any label): AGREEMENT if $\kappa \geq 0.60$ on every arm with at least 10 non-*other* rows and each such arm's human-W interval contains the instrument's W on the same rows; otherwise DISAGREEMENT, reported as an instrument-error bound with every disagreeing row listed and no rate corrected. Containment holds on 10 of 10 arms; $\kappa \geq 0.60$ on 9 of 10, failing on the

Qwen3-4B base checkpoint, so the registered verdict is DISAGREEMENT and the bound is stated from reader 1's read. Arm identities were hidden, but the base checkpoints' JSON wrapper makes their rows recognisable, so the read is arm-blind only for the corpus-format arms. Reader agreement on whitewash against refusal-reported, rows neither reader marked *other*: 0.968 raw, κ = 0.935 on 217 rows.

Table B.5: The two human reads against the automatic instrument on the same rows (template-group clustered 95% bootstrap, B = 2,000). The human W is the mean over template groups of each group's whitewash rate; the instrument is printed pooled over rows (the column the rule reads) and with the reader's estimator.

arm	rows read	other	reader 1 human W [CI]	instrument W, same rows	instrument W, reader's estimator	reader 1 κ	reader 1 disagreements	reader 2 human W [CI]	reader 2 κ	reader 2 disagreements	reader 1–reader 2 κ
released 35B flagship	20	0	0.18 [0.00, 0.35]	0.25	0.21	0.86	1	0.24 [0.06, 0.47]	0.73	2	0.86
dose-2	20	0	0.59 [0.38, 0.81]	0.50	0.59	1.00	0	0.59 [0.38, 0.81]	1.00	0	1.00
dose-v2	40	0	0.11 [0.02, 0.21]	0.15	0.10	0.80	2	0.12 [0.03, 0.23]	0.91	1	0.91
Qwen3-1.7B WM	20	0	0.64 [0.42, 0.83]	0.60	0.58	0.89	1	0.58 [0.36, 0.81]	1.00	0	0.89
Qwen3-1.7B base	20	0	0.94 [0.81, 1.00]	0.95	0.94	1.00	0	0.93 [0.80, 1.00]	1.00	0	1.00
Qwen3-4B WM	20	0	0.68 [0.47, 0.88]	0.60	0.62	0.89	1	0.59 [0.34, 0.81]	1.00	0	0.89
Qwen3-4B base	20	0	0.56 [0.33, 0.78]	0.70	0.72	0.40	6	0.50 [0.28, 0.72]	0.52	5	0.90
Qwen3-8B WM	20	0	0.24 [0.06, 0.44]	0.25	0.24	0.73	2	0.18 [0.03, 0.35]	0.86	1	0.86
Qwen3-8B base	20	0	0.38 [0.18, 0.62]	0.45	0.47	0.79	2	0.38 [0.18, 0.62]	0.79	2	1.00
SmolLM2-1.7B WM	20	0	0.69 [0.47, 0.88]	0.65	0.66	0.89	1	0.66 [0.44, 0.86]	1.00	0	0.89

On the corpus-format arms the disagreements are of three kinds: refusals the instrument missed (no error head, or a wrapper it does not parse: R011, R199, R208), a Y false positive on normal output (R177), and readers departing from the sheet's rule on a plain error string; every disagreement on the base checkpoints is a refusal inside the JSON wrapper that B.3's detector recovers. The agent's blind read of the same sheet, reported as a second automatic instrument and not as a human read, agrees with R at κ = 1.00 / 1.00 / 0.86 / 1.00 on the 1.7B / 4B / 8B / SmolLM2 arms and κ = 0.77 on dose-v2.

The flagship's own-benchmark sheet. The registered human read of the released model on its own benchmark (Appendix C.4) uses the primary tier less its two precondition/history rows (whose class has no counterpart among the paper's four), one prediction sample per row drawn by `random.Random(20260907)`, the A/B/C/X labels of the three-way protocol (A: the same refusal the environment produced; B: a different failure or refusal; C: a success or normal observation where the environment refused; X: unreadable), and shows the reader neither the instrument's verdict nor either agent reader's label; two rows produced no tagged sample, so the sheet holds 46 items over 40 trajectories. The registered bar was raw agreement ≥ 0.80 with agent reader 1 on the binary the paper uses: reader 1 0.978 (κ 0.950), reader 2 0.978 (κ 0.950), both validated. The human success-narration rate (trajectory-clustered bootstrap): reader 1 0.326 [0.196, 0.476], reader 2 0.326 [0.188, 0.477]; the paper prints their mean, 0.326, against agent reader 1's 0.304 on the same samples. Reader agreement on the binary: 0.957, κ = 0.901. The two disagreements with the agent read, one per reader, are printed below and were not adjudicated.

Table B.6: Every row on which a reader and the instrument disagree (reader 1: 16 rows; reader 2: 11; both readers give the same label on every shared row), with the prediction's head quoted. *wrapper*: the base checkpoints' JSON error wrapper the frozen head rule does not parse; *plain*: an error string at the head of the prediction, which the sheet's instructions label refusal-reported — on those rows the reader's label departs from the sheet's own rule and is kept as read.

row	arm	class	reader(s)	reader label	instrument	shape	prediction (head)
R011	Qwen3-8B WM	nonexistent-entity	reader 1, reader 2	REFUSAL-REPORTED	WHITE-WASH	other	<tool_response> Errno 2 No such file or directory [The command completed with return
R017	Qwen3-8B WM	unavailable-tool	reader 1	WHITE-WASH	REFUSAL-REPORTED	plain	<tool_response> {"error": "Unsupported tool: find_in_ text"} </tool_response>
R058	Qwen3-4B WM	unavailable-tool	reader 1	WHITE-WASH	REFUSAL-REPORTED	plain	<tool_response> {"error": "ERROR: Invalid tool name: find_in_text"} </tool_ response>
R077	Qwen3-1.7B WM	unavailable-tool	reader 1	WHITE-WASH	REFUSAL-REPORTED	plain	<tool_response> {"query": "2021-2023 ladder", "max_ results": 20, "error": "Unsupport
R090	SmolLM2-1.7B WM	unavailable-tool	reader 1	WHITE-WASH	REFUSAL-REPORTED	plain	<tool_response> {"query": "balance", "error": "Unsupported tool: find_in_ page"} </to
R106	Qwen3-4B base	unavailable-tool	reader 1, reader 2	REFUSAL-REPORTED	WHITE-WASH	wrapper	{"type": "text", "text": "Error calling tool 'find_in_ page': Unsupported tool: find_
R121	Qwen3-4B base	unmet-precondition	reader 1, reader 2	REFUSAL-REPORTED	WHITE-WASH	wrapper	{"type": "error", "message": "Payment method violation: Reservation can use at most
R130	Qwen3-8B base	nonexistent-entity	reader 1, reader 2	REFUSAL-REPORTED	WHITE-WASH	wrapper	{"type": "text", "text": "Error calling tool 'echo': The file 'test.txt' does not ex
R131	Qwen3-4B base	invalid-args	reader 1, reader 2	REFUSAL-REPORTED	WHITE-WASH	wrapper	{"type":"text","text":"Error calling tool 'filesystem-read_ file': 1 validation error
R142	Qwen3-4B base	unmet-precondition	reader 1, reader 2	REFUSAL-REPORTED	WHITE-WASH	wrapper	{"type": "text", "text": "Error placing buy order for ZETA: Insufficient funds. Requ
R151	Qwen3-8B base	nonexistent-entity	reader 1, reader 2	REFUSAL-REPORTED	WHITE-WASH	wrapper	{"type":"text","text":"Error calling tool 'cd': The folder 'document' does not exist
R153	Qwen3-4B base	nonexistent-entity	reader 1	WHITE-WASH	REFUSAL-REPORTED	wrapper	{"type":"text","text":"Error: Parent directory does not exist for path '/workspace/d
R160	Qwen3-4B base	unmet-precondition	reader 1, reader 2	REFUSAL-REPORTED	WHITE-WASH	wrapper	{"type": "error", "message": "Insufficient funds in account. Required funds for 100
R177	dose-v2	invalid-args	reader 1, reader 2	WHITE-WASH	REFUSAL-REPORTED	other	<tool_response> The file /testbed/reproduce_issue.py has been edited. Here's the res
R199	dose-v2	invalid-args	reader 1	REFUSAL-REPORTED	WHITE-WASH	other	<tool_response> find: path not specified. Ignoring /testbed/AUTHORS.md find: path no
R203	released 35B flagship	nonexistent-entity	reader 2	WHITE-WASH	REFUSAL-REPORTED	other	[{"type": "text", "text": "bash: tree: command not found\n", "annotations": null}, {
R208	released 35B flagship	unmet-precondition	reader 1, reader 2	REFUSAL-REPORTED	WHITE-WASH	other	{"status": "error", "message": "Cannot cancel order 12345: order is in a finalized s

Table B.7: Every item on which a reader and agent reader 1 disagree on the success-narration binary, with the prediction quoted. Labels: A the same refusal, B a different failure, C a success or normal observation, X unreadable.

item	reader	class	human	agent 1	prediction (head)
13	reader 1	interface/policy	C	B	`{"type":"text","text":"{\"message\":\"Reference does not exist\",\"documentation_ url\":\"...\",\"status\":\"404\"}"}`
38	reader 2	precondition/target	C	X	`[Old tool result content cleared]`

C FULL RESULTS FOR §3

C.1 WHITEWASH ON THE NATURAL ROWS, PER ARM AND PER CLASS

Supports Table 1 and Figure 1. Table C.1 prints the registered single run of every arm; the seed means of Table 1 are in Table D.2, and Figure C.1 plots every arm beside the base checkpoints. The 1.7B and 4B bands ("absent" ≥ 0.50, "partial" 0.20–0.50) are assigned on the point estimate, as registered; both intervals straddle 0.50.

Table C.1: Whitewash on the 399-row natural refusal set, per arm: macro-W with its template-clustered 95% CI, pooled W, and the per-class values. The flagship's row is read on 393 rows / 1,134 tagged samples; other arms on 398 rows (the 8B base arm 395) and SmolLM2 on 361. ‡: the base checkpoints and the flagship report some refusals inside a JSON wrapper the frozen rules do not parse; the wrapper-opened reading (B.3) is printed beside the registered one. †: the flagship's own-benchmark row is trajectory-clustered and not a macro-W.

Model	macro-W (whitewash) [95% CI]	pooled W (row-level mean) [95% CI]	per class: nonexistent / unmet / unavailable / invalid-args
Qwen3-1.7B	**0.525** [0.456, 0.604]	0.535 [0.463, 0.600]	0.545 / 0.579 / 0.426 / 0.549
Qwen3-4B	**0.456** [0.385, 0.537]	0.474 [0.415, 0.537]	0.502 / 0.524 / 0.361 / 0.436
Qwen3-8B	**0.263** [0.204, 0.336]	0.255 [0.207, 0.307]	0.236 / 0.271 / 0.306 / 0.240
SmolLM2-1.7B	**0.676** [0.600, 0.743]	0.694 [0.633, 0.754]	0.731 / 0.659 / 0.594 / 0.720
1.7B + dose 2	**0.607** [0.541, 0.671]	0.617 [0.560, 0.676]	0.633 / 0.656 / 0.579 / 0.559
1.7B + dose 1	**0.599** [0.536, 0.663]	0.617 [0.559, 0.675]	0.648 / 0.652 / 0.519 / 0.578
1.7B + dose-v2 (verbatim executor refusals)	**0.181** [0.120, 0.255]	0.195 [0.142, 0.261]	0.230 / 0.168 / 0.109 / 0.216
AgentWorld-35B-A3B, own format, **our rows**	**0.253** [0.183, 0.327]; its error wrapper counted‡: 0.22 [0.15, 0.30]	0.274 [0.210, 0.348]	0.332 / 0.208 / 0.137 / 0.333 (‡ unavailable 0.044)
Qwen3-1.7B / 4B / 8B **base, no WM SFT**, flagship's contract	rules: **0.760** [0.697, 0.814] / **0.647** [0.580, 0.726] / **0.516** [0.435, 0.606]; their own error wrapper counted‡: **0.68** [0.60, 0.75] / **0.44** [0.36, 0.54] / **0.41** [0.33, 0.51]	0.765 / 0.659 / 0.537	paired ΔW = SFT – base‡: –0.14 [–0.22, –0.06] / +0.03 [–0.03, +0.09] / –0.18 [–0.25, –0.12] (rules alone: –0.23 / –0.19 / –0.28; per class Appendix C.3)
AgentWorld-35B-A3B, own format, **its own benchmark**†	— (not a macro-W)	—	W = **0.39**† [0.26, 0.53], trajectory-clustered; three-way read: 0.33 of samples narrate a success

Table C.2: The prompted arm (Qwen3-30B-A3B-Instruct-2507, bf16, no training, the flagship's prompt contract) against the released flagship, per class, same rows, same instrument.

class	prompted Qwen3-30B-A3B	flagship
nonexistent-entity	0.485	0.332
unmet-precondition	0.443	0.208
invalid-args	0.613	0.333
unavailable-tool	0.448	0.137

Two analysis-only splits. Rows whose refusal template already occurs in the prompt's own history (35 of 399) are reproduced by every model: whitewash 0.12 / 0.16 / 0.20 at 1.7B / 4B / 8B against 0.57 / 0.50 / 0.26 on the other rows (at 8B the split is inside noise). Bucketing rows instead by

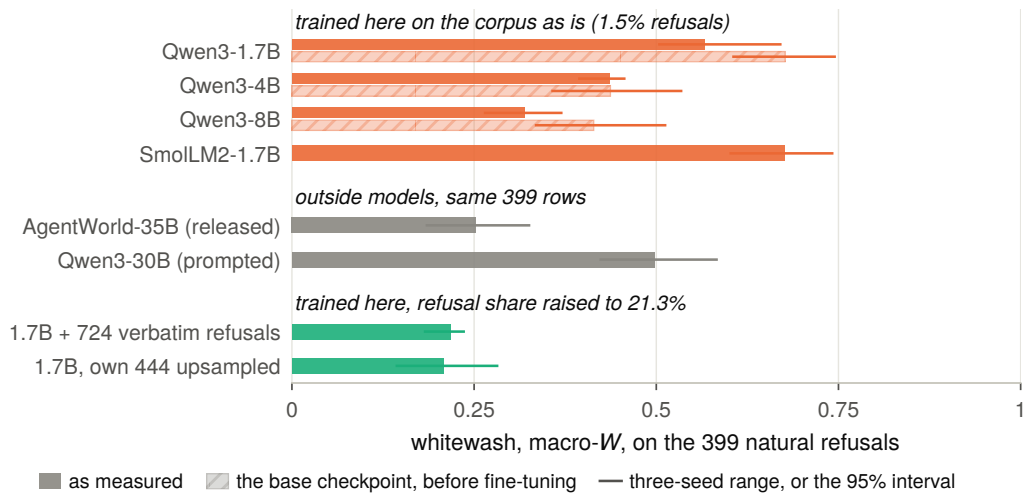


Figure C.1: Whitewash on the 399 natural refusals for every arm of Table 1. Solid bars are the arm as measured — the seed mean where it has three seeds, otherwise the single registered run. Hatched bars are the Qwen3 base checkpoints before fine-tuning, read with the JSON error wrapper opened (B.3). Lines are the three-seed range, or the 95% template-clustered interval where the arm is one run.

how many *training* rows share their refusal template (0 / 1–4 / ≥ 5: 82 / 111 / 206 rows) moves nothing: the 8B reports the refusal on 0.76 of rows whose template it never saw in training (0.243 [0.15, 0.35] seen-0$\times$ against 0.269 [0.19, 0.36] seen-$\geq$5$\times$), so the larger models' recovery is not a training-template lookup. **Domain.** SWE whitewash exceeds API whitewash in every arm's point estimate — 1.7B 0.81 [0.67, 0.94] against 0.51 [0.45, 0.57]; 4B 0.74 [0.48, 0.92] against 0.45 [0.39, 0.51]; 8B 0.48 [0.19, 0.78] against 0.24 [0.19, 0.28]; SmolLM2 0.85 [0.60, 1.00] against 0.68 [0.62, 0.74] — on 32 SWE rows against 367 API rows, so the intervals separate only at 1.7B.

The blanket-refuser tripwire. The registered rule marks a blanket refuser when $R > 0.30$ on the legal companion under the full union, whose template term on a legal row detects fidelity rather than refusal and alone contributes 0.24–0.31. Under the literal union the rule fires for six of the seven SFT arms; under the two terms that can detect a refusal on a normal row (head $\vee$ Y) the rate is 0.04–0.11 and it fires for none. The paper reports head $\vee$ Y and records this in Table G.1 as the only place a registered rule is set aside; a reader who holds us to the literal registration should treat those arms' W as unquotable and read boundary recovery on the natural rows as 0.20 / 0.33 / 0.61 / 0.07 / 0.11 / 0.005 / 0.69 (1.7B / 4B / 8B / dose 2 / dose 1 / SmolLM2 / dose-v2) instead of the 0.43 / 0.51 / 0.73 / 0.34 / 0.33 / 0.26 / 0.78 under head $\vee$ Y.

C.2 The planted battery, every cell

Supports Table 2 and §3.2.

What the Y-harvested doses moved was the listing share: `cat -n` emissions rose from 38% of the battery's decorated-register emissions at base to 78% (dose 1) and 67% (dose 2), and Y labels 22–56% of such listings `error`. The class-consistent share — refusals that name what was planted — stays at 0.000–0.013 on every arm.

C.3 The base checkpoints

Supports §3.1 (the base checkpoints). The Qwen3 base checkpoints, prompted under the flagship's `<predicted_observation>` contract on the same 399 rows (tag fraction 1.000 / 0.994 / 0.936 against the 0.85 gate), whitewash 0.760 / 0.647 / 0.516 under the registered rules; per class 0.766 / 0.810 / 0.749 / 0.716, 0.667 / 0.725 / 0.536 / 0.662, 0.600 / 0.453 / 0.456 / 0.554 (nonexistent / unmet / unavailable / invalid-args); legal-row refusal (head $\vee$ Y) 0.036 / 0.038 / 0.037. The human read (B.5) showed the rules under-read these arms: every disagreement is a refusal reported inside a JSON

Table C.3: Matched boundary recovery on the battery under the head rule: $(p_{\text{head}}(\text{invalid}) - p_{\text{head}}(\text{legal}))/(1 - p_{\text{head}}(\text{legal}))$ on the class's own states, state-clustered 95% CIs (B = 2,000, paired resampling of states); exact-zero cells have degenerate intervals. dose 1 / dose 2 are the Y-harvested arms of B.2 and dose-v2 the verbatim-refusal repair of §4.1.

Class	1.7B	4B	8B	SmolLM2	dose 1	dose 2	dose-v2
nonexistent-entity	0.000	0.009 [0.00, 0.03]	0.028 [0.00, 0.08]	0.000	0.005 [–0.02, 0.03]	0.014 [0.00, 0.04]	0.000
unmet-precondition	0.000	0.036 [–0.01, 0.10]	0.042 [0.005, 0.097]	0.000	**0.188** [0.12, 0.26]	0.084 [0.04, 0.14]	0.088 [0.02, 0.17]
unavailable-tool	0.027 [0.01, 0.04]	0.009 [0.00, 0.02]	–0.002 [–0.01, 0.00]	0.014 [0.00, 0.03]	0.074 [0.05, 0.10]	0.044 [0.02, 0.07]	0.079 [0.05, 0.11]
invalid-args	0.004 [0.00, 0.01]	0.045 [0.01, 0.08]	**0.190** [0.11, 0.29]	0.000	0.027 [0.00, 0.05]	0.027 [0.004, 0.05]	0.088 [0.02, 0.17]
matched baseline per class (nonexistent / unmet / unavailable / invalid-args)	0.000 / 0.000 / 0.008 / 0.000	0.000 / 0.004 / 0.009 / 0.004	0.023 / 0.040 / 0.017 / 0.040	0.000 / 0.000 / 0.011 / 0.000	0.009 / 0.009 / 0.011 / 0.009	0.000 / 0.000 / 0.012 / 0.000	0.023 / 0.044 / 0.017 / 0.044
matched macro-BR	0.008 [0.00, 0.01]	0.025 [0.01, 0.04]	0.064 [0.04, 0.09]	0.004 [0.00, 0.01]	0.074 [0.05, 0.09]	0.042 [0.03, 0.06]	0.064 [0.04, 0.09]
unmatched macro-BR (all-states baseline)	0.002 [–0.002, 0.007]	0.020 [0.00, 0.04]	0.076 [0.04, 0.11]	–0.005 [–0.010, 0.001]	0.072 [0.05, 0.09]	0.033 [0.02, 0.05]	0.078 [0.04, 0.12]

Table C.4: The same cells read with the classifier Y instead of the head rule (same emissions, states, matched baselines and seed).

Class	Qwen3-1.7B	Qwen3-4B	Qwen3-8B	SmolLM2-1.7B
nonexistent-entity (state)	0.005 [–0.03, 0.04]	0.078 [0.01, 0.16]	0.035 [–0.01, 0.10]	-0.008 [–0.04, 0.02]
unmet-precondition (state)	0.039 [–0.02, 0.09]	0.048 [–0.02, 0.13]	0.075 [0.01, 0.14]	0.024 [–0.04, 0.09]
unavailable-tool (static)	0.025 [–0.01, 0.06]	-0.003 [–0.04, 0.03]	0.012 [–0.01, 0.04]	0.021 [–0.00, 0.04]
invalid-args (static)	-0.039 [–0.10, 0.00]	0.168 [0.08, 0.27]	0.165 [0.07, 0.27]	-0.015 [–0.06, 0.02]
matched macro-BR, Y	**0.007** [–0.02, 0.03]	**0.073** [0.04, 0.11]	**0.072** [0.04, 0.11]	**0.005** [–0.02, 0.03]

Table C.5: The Y-harvested repair read with the head rule: p_{head} per class with state-clustered 95% CIs, matched legal baselines in parentheses, and the class-consistent share (emissions that name the planted element).

Class	unmodified 1.7B	dose 1	dose 2
unmet-precondition (held out · state)	0.000 (0.000)	0.196 [0.13, 0.27] (0.009)	0.084 [0.04, 0.14] (0.000)
nonexistent-entity (held out · state)	0.000 (0.000)	0.014 [0.00, 0.03] (0.009)	0.014 [0.00, 0.04] (0.000)
unavailable-tool (trained)	0.035 [0.02, 0.05] (0.008)	0.084 [0.06, 0.11] (0.011)	0.055 [0.04, 0.07] (0.012)
invalid-args (trained)	0.004 [0.00, 0.01] (0.000)	0.036 [0.01, 0.06] (0.009)	0.027 [0.01, 0.05] (0.000)
legal baseline, all 199 states	*0.008*	*0.011*	*0.012*
matched macro-BR [95% CI]	**0.008** [0.00, 0.01]	**0.074** [0.05, 0.09]	**0.042** [0.03, 0.06]
class-consistent share (names the planted element; n = 44 / 45 / 199 / 45 states)	0.000–0.014	**0.000–0.004**	**0.000–0.013**

wrapper the head rule does not match, which the base arms emit on 8.5 / 21.3 / 9.5% of samples and the SFT arms on none. Counting the wrapper (B.3) gives 0.677 [0.604, 0.746] / 0.437 [0.356, 0.536] / 0.414 [0.333, 0.514] and paired ΔW (SFT – base) of –0.144 [–0.221, –0.062] / +0.028 [–0.032, +0.089] / –0.181 [–0.253, –0.116]; under the rules alone the deltas are –0.229 / –0.185 / –0.280, every per-class CI below 0. The paper's sentences use the wrapper-opened reading; the registered numbers stand as an upper bound on the base arms' W (Table G.1). The two members of each pair run under different contracts (the training format against the flagship's tag instruction).

C.4 The released flagship on its own benchmark

Supports §3.1 (the released flagship) and §6. **Selection, which was not a sample.** AgentWorld-Bench ships 1,905 held-out trajectories. We scanned every turn whose recorded observation is an environment refusal (2,170 candidate rows), read and classed every candidate before any prediction was scored (117 rows read in full; the remaining Y-only hits are format false positives such as page snapshots carrying "Console: N errors"), and froze a 60-row set: 50 primary rows — precondition/target 16, precondition/tool 18 (the *unavailable-tool* shape), interface/schema 9, interface/policy 5, precondition/history 2 (no counterpart among the paper's four classes) — and 10 program-level rows kept as a secondary tier. The other 57 read rows are excluded by class and listed with the verdicts: 40 are not refusals (a harness printing `Error: (none)` on exit 0, a `cat` of a file that contains a traceback, `grep` exiting 1), 11 are infrastructure failures (API 400/422, MCP parse failures, browser network errors), 3 are a human declining the call, 2 hide the malformed agent action from the world model, 1 is an output-cap refusal. All 60 were prompted in the benchmark's own assembly (system prompt, full prior history, the `<predicted_observation>` contract), five predictions per row, 16,384-token budget; 57 were scored (three rows emitted 0 of 5 tagged predictions — the model reasons past the budget without opening the tag), 48 of the 50 primary. Every own-benchmark number is over those 48 rows, clustered over their 41 trajectories, 225 of 240 samples tagged.

Readings. Registered whitewash (the instrument) 0.39 [0.26, 0.53]; three-way read (reproduces the refusal / other failure / narrates a success) 0.33 [0.20, 0.47] of samples narrate a success, validated by two human readers at 0.326 (B.5). The conditioning is the finding: the flagship reproduces every interface/schema refusal (6/6 trajectories, 36/36 samples), 0.71 of precondition/target refusals and 0.27 of precondition/tool refusals; rows whose refusal line already occurs earlier in the same trajectory are reproduced on 0.90 [0.71, 1.00] (12 trajectories) and novel rows on 0.45 [0.30, 0.62] (34). The aggregate sits where the vendor's own "Error Handling" axis sits — 60–69% accuracy on its text domains (MCP 68.9, Terminal 67.3, SWE 60.1), an axis that checks deterministic-error turns for an error signal rather than a plausible success, is not in the headline rubric, selects turns by outcome rather than by action validity, and sits downstream of a pipeline that removes the error-and-retry turns whose absence §1 diagnoses. **Quantisation.** The released weights are AWQ-INT4. Re-served at full precision on the 399 rows, macro-W moves from 0.250 to 0.255 (paired over 160 template groups: +0.007 [–0.010, +0.022]; quantisation is not the cause); on the 48 own-benchmark rows the full-precision arm reports no refusal on 0.372 against INT4's 0.391; its validity-gate score is 0.759 against INT4's 0.774, both below the 0.80 bar.

On our rows. In its own contract on the 399 rows and 185 companions (1,134 of 1,194 refusal samples tagged, 0.950): macro-W 0.253 [0.183, 0.327], pooled 0.274 [0.210, 0.348]; legal false refusals 0.031 [0.007, 0.064]. A third wrapper family (`{"status": "error", ...}`) appears in 1.9% of its tagged samples; counting it gives 0.224 [0.153, 0.300].

C.5 What whitewash is: success narration

Supports §3.1 (success narration). Every whitewashed 8B sample on the natural rows (305 samples over 139 rows) was read by the three-way protocol of C.4 — three blind agent reads, two runs of one model agreeing on 0.987 of samples and a third model agreeing with the primary on 305/305 — as analysis only, with the registered W unchanged by any label: 0.21 [0.16, 0.26] of all 1,194 samples narrate a success or a normal observation where the executor refused (0.83 of the whitewash), against 0.01 [0.01, 0.02] that are the executor's own refusal missed by R, 0.01 a different failure, and 0.02 unreadable (a bare `None`, an argument echo). The human calibration of the same instrument is Table B.5.

C.6 THE DECIDABILITY READ

Supports §3.1 (the decidable rows), the decidable column of Table 1, and §7. **The question.** Whether the prefix shown to the simulator holds the fact that decides the executor's refusal — a file named as present that the listing lacks, an order whose status forbids the intent, a tool the state never offers — or whether the refusal could only be learned by running the action. A row is DECIDABLE when a reader can quote a span of the shown prefix from which the refusal follows, UNDECIDABLE when the prefix holds no such span, UNCLEAR when the reader cannot tell; a DECIDABLE label with no quotable span counts as UNCLEAR, and every quoted span is checked mechanically to occur in the prefix that reader saw. **The reads.** The assisting language model labelled all 399 rows under the registered protocol (each row shown only the system prompt, the state prefix, the recorded action and the recorded observation; no arm's emission, no readout), in one pass whose model version changed once, at row 360, the change recorded before those rows were labelled. Because the recorded observation is shown, both reads judge whether the refusal is explicable from the prefix once the outcome is known, a weaker standard than predicting it from the prefix alone; the bias runs toward DECIDABLE, so the decidable share is an upper bound. An author labelled a stratified sample of 100 rows (25 per class, chosen by a seeded hash of the row id, the list written before the model's read began), in a registered shuffled order with no row skipped, never shown the model's labels. A first read of the same 100 rows, by a reader outside the author list and before the span requirement existed, was withdrawn on a protocol ground before any reading was computed on it: it recorded no spans, so under the registered rule every one of its DECIDABLE labels counts as UNCLEAR. The second read, by an author under the registered protocol, replaced it in full; both are in the supplementary material side by side. The row file of the prefix counterfactual (D.6) still pins the withdrawn read's labels, and those rows were used as committed. **The gate.** The model's labels stand in for a human read on the remaining 299 rows only if they agree with the author's read on at least 0.80 of the 100. They agree on 85 ($\kappa = 0.694$ over three categories): of the author's 58 DECIDABLE rows the model calls 51 decidable, 6 undecidable and 1 unclear; of the 42 UNDECIDABLE rows, 34 undecidable and 8 decidable. So the stratum is the model's labels on all 399 rows: 230 DECIDABLE, 168 UNDECIDABLE, 1 UNCLEAR (excluded and counted). **The reading.** Within each class holding at least 5 rows in both strata, W_{dec} and W_{undec} are the mean over rows of the row's W; the headline is the macro over those classes of $W_{\text{undec}} - W_{\text{dec}}$, so the class mix is held fixed, with a template-clustered bootstrap (groups resampled within class and stratum, B = 2,000, one pick-list shared across a model's seeds). Branches fixed before any label: at least 0.15 with the interval excluding zero, the deficit sits where the outcome must be inferred from the state; below 0.10 in magnitude, decidability does not carry it; between, or the interval containing zero, printed as read. For the 1.7B seed mean the contrast is **+0.117 [+0.002, +0.223]**: the third branch. The interval reaches down to 0.00, and the text leans on the two absolute values, 0.54 on the decidable rows against 0.66 on the undecidable ones, not on that edge. Three classes qualify (Table C.8); unavailable-tool does not, because a tool absent from the tool list is decided by the list itself, leaving 2 undecidable rows. The three do not agree in direction — invalid-args +0.43, nonexistent-entity +0.03, unmet-precondition –0.10 — so the macro is carried by invalid-args, which is what §3.1 says. Every arm is in Table C.7; the released flagship, at 0.14 against 0.48, is the arm whose deficit is most concentrated on the undecidable rows. The withdrawn first read, on its own 100 rows with one qualifying class, would have put the 1.7B contrast at +0.36 [+0.19, +0.59] and fired the first branch; the two readings are printed side by side in the supplementary material. **Copyability, not decidability** (registered, descriptive): among the 168 undecidable rows, 7 carry their refusal template earlier in the prefix, and on those the 1.7B whitewashes 0.14 against 0.61 on the other 161 — seven rows, listed because the reading was registered. **Provenance.** The model's 399 labels were frozen on a dated commit eight days before the author's read landed, their hash pinned by the readout; the reading's rules, floors and branches were registered before either read. The withdrawn read's labels selected the rows of the prefix counterfactual (D.6), a selection kept as committed; no number reported in this paper is computed from those labels.

One class set for every stratum, and the labels' error. Table 1's all-rows column is a macro over four classes and its decidable column over the three that qualify, so the two differ by the class set as well as by the stratum. Table C.9 puts the three strata on the same three classes and the same row-level weighting, with the rows each arm scores; the ordering is unchanged for every arm. The second half reads the 100 rows the author labelled by agreement between the two readers: on the 51 rows both call decidable, the 34 both call undecidable and the 15 on which they differ. Label error moves the decidable estimate by a few hundredths (the 1.7B seed mean 0.54 on the reader's

Table C.7: Whitewash by decidability stratum, every arm: macro-W over the three qualifying classes on the 230 decidable and the 168 undecidable rows, with the registered contrast $W_{\text{undec}} - W_{\text{dec}}$ and its 95% interval. Seed means where the arm has three seeds.

arm	W, decidable rows	W, undecidable rows	undecidable – decidable
Qwen3-1.7B	0.544 [0.484, 0.619]	0.661 [0.571, 0.749]	+0.117 [+0.002, +0.223]
Qwen3-4B	0.400 [0.330, 0.489]	0.587 [0.455, 0.690]	+0.187 [+0.034, +0.304]
Qwen3-8B	0.239 [0.180, 0.308]	0.409 [0.283, 0.532]	+0.170 [+0.024, +0.298]
SmolLM2-1.7B	0.643 [0.568, 0.731]	0.775 [0.671, 0.849]	+0.131 [–0.004, +0.234]
1.7B + dose 1	0.567 [0.503, 0.644]	0.692 [0.592, 0.763]	+0.125 [+0.001, +0.218]
1.7B + dose 2	0.554 [0.482, 0.641]	0.695 [0.626, 0.765]	+0.141 [+0.025, +0.243]
1.7B + dose-v2 (verbatim executor refusals)	0.156 [0.108, 0.202]	0.268 [0.105, 0.399]	+0.113 [–0.057, +0.251]
AgentWorld-35B-A3B, own format (INT4, as released)	0.137 [0.090, 0.200]	0.480 [0.367, 0.612]	+0.344 [+0.213, +0.480]
AgentWorld-35B-A3B, own format (bf16)	0.157 [0.105, 0.218]	0.475 [0.364, 0.595]	+0.318 [+0.189, +0.449]
Qwen3-30B-A3B-Instruct, prompted	0.359 [0.275, 0.469]	0.695 [0.614, 0.789]	+0.336 [+0.192, +0.458]

Table C.8: The four classes by stratum: rows scored in each stratum (the same for every full-coverage arm; one unmet-precondition decidable row has no tagged sample, so the reader's 230 decidable rows score as 229), and the 1.7B seed mean's W in each. A class qualifies for the macro when both strata hold at least 5 rows.

class	decidable rows	undecidable rows	W_{dec}	W_{undec}	undecidable – decidable	in the macro
nonexistent-entity	77	101	0.600	0.626	+0.026	yes
unmet-precondition	56	35	0.655	0.553	–0.102	yes
invalid-args	38	30	0.377	0.804	+0.427	yes
unavailable-tool	58	2	0.460	1.000	+0.540	no (2 rows)

decidable rows against 0.48 where both readers agree; the released model 0.14 against 0.09) and does not change which stratum whitewashes more. Both reads saw the recorded refusal (C.6), so the sensitivity is to disagreement between readers, not to that shared bias.

Table C.9: Top: macro-W over the three qualifying classes on every row of those classes, on the decidable rows and on the undecidable rows, with the rows scored (n) and template-clustered 95% intervals; seed means where the arm has three seeds. Bottom: W on the author's 100 rows by agreement between the author's read and the LLM reader's, macro over the classes present (the 51-row and 15-row subsets hold all four classes; the 34-row subset three).

arm	all rows, three classes	n	decidable	n	undecidable	n
Qwen3-1.7B	0.599 [0.520, 0.674]	337	0.544 [0.484, 0.619]	171	0.661 [0.571, 0.749]	166
Qwen3-4B	0.482 [0.395, 0.565]	337	0.400 [0.330, 0.489]	171	0.587 [0.455, 0.690]	166
Qwen3-8B	0.314 [0.239, 0.401]	337	0.239 [0.180, 0.308]	171	0.409 [0.283, 0.532]	166
SmolLM2-1.7B	0.703 [0.618, 0.770]	315	0.643 [0.568, 0.731]	160	0.775 [0.671, 0.849]	155
1.7B + dose 1	0.626 [0.546, 0.695]	337	0.567 [0.503, 0.644]	171	0.692 [0.592, 0.763]	166
1.7B + dose 2	0.616 [0.534, 0.690]	337	0.554 [0.482, 0.641]	171	0.695 [0.626, 0.765]	166
1.7B + dose-v2 (verbatim executor refusals)	0.205 [0.127, 0.292]	337	0.156 [0.108, 0.202]	171	0.268 [0.105, 0.399]	166
AgentWorld-35B-A3B, own format (INT4, as released)	0.291 [0.203, 0.381]	332	0.137 [0.090, 0.200]	169	0.480 [0.367, 0.612]	163
AgentWorld-35B-A3B, own format (bf16)	0.298 [0.208, 0.389]	331	0.157 [0.105, 0.218]	170	0.475 [0.364, 0.595]	161
Qwen3-30B-A3B-Instruct, prompted	0.514 [0.421, 0.608]	337	0.359 [0.275, 0.469]	171	0.695 [0.614, 0.789]	166
the author's 100 rows	*both decidable (51)*		*both undecidable (34)*		*readers differ (15)*	
Qwen3-1.7B	0.479 [0.358, 0.613]	51	0.628 [0.477, 0.728]	34	0.634 [0.565, 0.717]	15
Qwen3-8B	0.219 [0.134, 0.320]	51	0.422 [0.245, 0.544]	34	0.361 [0.244, 0.469]	15
AgentWorld-35B-A3B (INT4, as released)	0.092 [0.021, 0.202]	51	0.479 [0.263, 0.668]	31	0.208 [0.107, 0.381]	14

D FULL RESULTS FOR §4

D.1 THE REFUSAL-SHARE LADDER AND ITS SEEDS

Supports §4.1 and Figure 2a–b. The dose-v2 recipe, byte-identical, at aug-ratio 20, 10 and 4 between the arms already on record, so that W and the probe AUC are read at five refusal shares of the training mixture (share = head-rule rows over all training rows; every rung ≥ 2 sees the whole 724-row bank, about $2\times$ / $4\times$ / $10\times$ per row; one training seed per rung). Two statements were fixed before the

data: monotone emission (macro-W non-increasing in share: yes) and flat representation (probe AUC within ± 0.05 of rung 1's 0.805 at every rung: yes, –0.038 at worst). Row counts from the train logs: rung 0 removes the 444 head-rule rows (29,113 remain); rungs 2–4 add 1,477 / 2,955 / 7,387 bank rows.

Table D.1: The ladder on the 399 natural rows: macro-W with its template-clustered CI, paired ΔW against rung 1 over the rows both arms scored, legal-row false refusals (head $\vee$ Y), the probe's natural-row AUC, the emission's AUC and the probe-minus-emission gap (group-clustered bootstrap, B = 2,000).

rung	share	macro-W [CI]	ΔW vs rung 1 [CI]	R legal	probe AUC [CI]	E-AUC [CI]	gap [CI]
0 — the 444 head-rule rows removed	0.0%	0.670 [0.608, 0.735]	+0.153 [+0.115, +0.193]	0.040	0.766 [0.688, 0.833]	0.519 [0.427, 0.601]	+0.247 [+0.142, +0.356]
1 — the 1.7B world model (slice's own rows)	1.5%	0.525 [0.456, 0.604]	—	0.067	0.805 [0.738, 0.865]	0.584 [0.494, 0.666]	+0.221 [+0.132, +0.315]
2 — dose-v2 recipe, aug-ratio 20	6.2%	0.290 [0.216, 0.381]	—	0.133	0.768 [0.693, 0.831]	0.694 [0.607, 0.772]	+0.074 [–0.019, +0.171]
3 — dose-v2 recipe, aug-ratio 10	10.5%	0.233 [0.166, 0.319]	—	0.130	0.797 [0.734, 0.854]	0.722 [0.641, 0.795]	+0.076 [–0.006, +0.168]
4 — dose-v2 (aug-ratio 4)	21.3%	0.181 [0.120, 0.255]	–0.340 [–0.395, –0.287]	0.110	0.802 [0.731, 0.867]	0.743 [0.663, 0.814]	+0.059 [–0.026, +0.151]

Power. Each rung is one training run and every interval in Table D.1 is over rows, so none carries run-to-run variation. Re-training the same 1.7B recipe with only the seed changed moves macro-W with an SD of 0.092 (Table D.2). The endpoints differ by 0.489 as single runs and by 0.453 at their seed means; the four interior steps are 0.145, 0.235, 0.056 and 0.052, and the last two are smaller than the seed SD. The 0.181 printed for rung 4 is that arm's first seed; its three-seed mean is 0.217 (range 0.181–0.237, SD 0.032).

Table D.2: Training-seed replicates of the three Qwen3 arms: macro-W on the 399 rows per seed (the seed drives the 30,000-row subsample and the trainer), mean, range, SD, the registered band, and the false-refusal rate on the 185 legal companions (head $\vee$ Y). The disjoint-range call has null probability 0.10 per pair at three against three seeds.

scale	seeds read	macro-W per seed	mean	range	seed SD	band(s)	legal FP, mean (range)
1.7B	3	0.525 / 0.672 / 0.503	0.566	0.503–0.672	0.092	boundary absent	0.050 (0.034–0.067)
4B	3	0.456 / 0.458 / 0.393	0.435	0.393–0.458	0.037	partial	0.042 (0.040–0.045)
8B	3	0.263 / 0.371 / 0.323	0.319	0.263–0.371	0.054	partial	0.036 (0.025–0.050)

D.2 Bank against share, and coverage against frequency

Supports §4.1 (*Bank or share?*). The ladder moves the refusal share and the augmentation bank together. Two arms fill the empty cells of the 2×2 at the shares already on record: **A** trains on the dose-v2 bank with the slice's own head-rule rows dropped, at a ratio landing near 1.5%; **B** upsamples the slice's own 444 rows to the share dose-v2 reaches. Both are read on the frozen 399 rows with the instrument unchanged; the realised share is read from the trainer's own row counts.

Table D.3: Macro-W in the bank $\times$ share 2×2. The diagonal is on record; the off-diagonal is the amendment.

bank / share	$\approx$1.5%	$\approx$21.3%
the slice's own rows	0.525 [0.456, 0.604]	**B** 0.208 [0.142, 0.283]
the dose-v2 bank	**A** 0.566 [0.493, 0.644]	0.181 [0.120, 0.255]

At fixed 1.5% the two banks do not separate (Δ +0.041, smaller than the 0.092 seed SD, and with three seeds a side — bank 0.566 / 0.525 / 0.533 against the slice's own 0.525 / 0.672 / 0.503 — the sets overlap); at fixed bank the share moves W (B against the reference: Δ –0.317, intervals disjoint). **Reach of the repair.** Split by the bank's own template groups (one run), W falls 0.40 $\rightarrow$ 0.07 on the 99 natural rows the bank covers and 0.56 $\rightarrow$ 0.21 on the 288 it never contained (11 further rows the bank covers by template but not by group are in neither set, and one outside row has no tagged

Table D.4: The same four arms with their realised share, the false-refusal rate on the 185 legal companions (head ∨ Y) and the samples scored.

arm	bank	realized share	macro-W [95% CI]	legal FP	samples scored
the 1.7B world model (reference)	the slice's own 444 rows	1.5%	0.525 [0.456, 0.604]	0.067	1194
A: the repair bank at 1.5%	the dose-v2 bank	1.5%	0.566 [0.493, 0.644]	0.052	1194
B: the slice's own rows upsampled	the slice's own 444 rows, upsampled	21.3%	0.208 [0.142, 0.283]	0.119	1194
the verbatim-refusal repair (reference)	the dose-v2 bank	21.3%	0.181 [0.120, 0.255]	0.110	1194

sample); on the battery, whose tools the bank excludes, the repair lifts matched recovery from 0.008 to 0.064 (Table C.3, dose-v2 column).

Coverage against frequency. Four arms cross the number of refusals with their spread over the tool support, the dose-v2 recipe byte-identical except the bank; row counts are equal within a share and not across it (29,554 at 1.5%, 36,391 at 20%), so the high-share arms also saw 23% more rows.

Table D.5: The coverage × frequency 2×2 on the registered test tiers (211 held-out-combination rows, 187 entirely-unseen-tool rows); macro-W under the registered union, template-clustered bootstrap B = 2,000; false refusal is head ∨ Y on the 185 legal companions.

arm	share	rows	cells	tools	W, held-out combination	W, unseen tool	W, all	legal false refusal
low / distributed	1.5%	441	113	64	0.537 [0.468, 0.616]	0.703 [0.594, 0.809]	0.576 [0.511, 0.652]	0.040
low / concentrated	1.5%	441	51	8	0.563 [0.485, 0.647]	0.735 [0.630, 0.829]	0.598 [0.535, 0.671]	0.047
high / concentrated	20.0%	7,278	51	8	0.378 [0.302, 0.469]	0.612 [0.477, 0.734]	0.425 [0.350, 0.516]	0.067
high / distributed	20.0%	7,278	113	64	0.236 [0.159, 0.293]	0.412 [0.234, 0.560]	0.259 [0.189, 0.339]	0.139

The arm with ten times the refusals is the lower one wherever they sit (0.537 against 0.378 on the held-out-combination rows), so coverage is not established as the governing factor; within a share, distributed against concentrated is disjoint at the high share (0.236 against 0.378) and overlapping at the low one (0.537 against 0.563, inside the seed SD); the low-W arm carries the highest false-refusal rate (0.139).

D.3 The legality control

Supports §4.2 (*Rarity*). A legal observation class is held at the refusal share (1.5%) and raised to 21.3%, read with the template term alone: U = 1 – (fraction of samples reproducing the class's observation), the legal analogue of W. Two classes: one action-inferable (`File created successfully at: ...`, 400 rows, 2.22% of the corpus) and one state-dependent (`... did not produce any output`, 61 rows, 0.54%). The checkpoint is a retrain of the 1.7B recipe; its pooled W on the 399 rows (0.664) lies inside the original run's stored interval.

Table D.6: U on the two legal-class row sets and W on the 399 refused rows from the same two checkpoints; U with its row-level interval, W with its template-clustered one.

reading	rung 1 (1.5%)	rung 4 (21.3%)	contrast	branch
U, action-inferable class (400 rows)	0.036 [0.022, 0.051]	0.007 [0.001, 0.015]	ΔU 0.029 against ΔW 0.026: matches	2: under-generated beyond rarity
U_b, state-dependent class (61 rows)	0.525 [0.410, 0.634]	0.454 [0.350, 0.568]	ΔU 0.071 against ΔW 0.026: matches	3: between
W on the 399, registered union	0.632 [0.569, 0.702]	0.606 [0.544, 0.674]	0.026	placebo: inside the seed range
legal-companion false refusals, head ∨ Y (185 rows)	0.040	0.029	—	descriptive

D.4 THE LINEAR PROBE

Supports §4.2 (*Representation*); Figure D.1 plots the probe per layer and per model. A standardised logistic regression (C = 1, balanced classes) on the residual-stream hidden state at the last prompt token — the position the observation is generated from — one probe per layer, layer chosen by nested cross-validation inside grouped five-fold CV over template groups (161 refusal groups, 53 legal groups; no group split across folds). The label-permutation control, the identical pipeline with labels shuffled, reads 0.47–0.52 for the three world models and 0.40–0.56 over the nine arms of Table D.8 (one seeded permutation per arm, so no interval is printed).

Table D.7: The executor's refusals against its executions on the 584 natural rows: the state (probe) against the same rows' emission read two ways — the refusal-only instrument at twenty samples per row, and a trained text read-out (TF-IDF word 1–2-grams with a balanced logistic regression, refit out of fold under the probe's own grouped five folds) — with the paired state-minus-emission gap. Both emission read-outs are post hoc controls, registered after they were computed.

read of the refusal — AUC, refused vs legal (584 rows; emission 583)	Qwen3-1.7B	Qwen3-4B	Qwen3-8B
state — probe on the residual stream	**0.81** [0.74, 0.86]	**0.84** [0.79, 0.89]	**0.94** [0.90, 0.96]
— the same probe, API-tool / SWE rows (536 / 48)	0.85 / 0.29	0.87 / 0.50	0.95 / 0.71
emission — refusal-only instrument, 20 samples/row	0.80	0.84	0.92
emission — trained text readout	0.63	0.67	0.83
state – emission (refusal-only, 20 samples)	+0.010 [–0.056, +0.070]	+0.000 [–0.046, +0.046]	+0.019 [–0.005, +0.043]

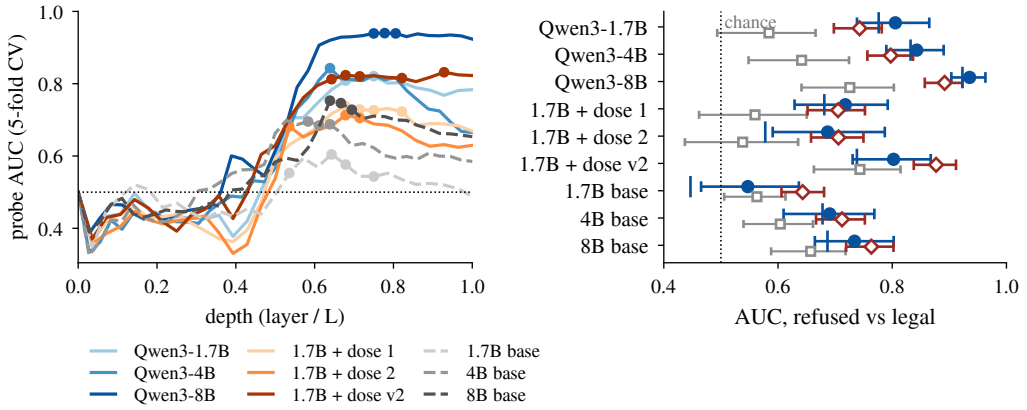


Figure D.1: Left: the per-layer grouped-CV AUC of the probe per model on a shared depth axis, the nested-choice layers marked. Right: the probe's AUC beside the emission's registered AUC with CIs, the within-tool AUC as a tick.

Per class, the probe's AUC (nonexistent-entity / unmet-precondition / unavailable-tool / invalid-args): 1.7B 0.819 / 0.765 / 0.864 / 0.772; 8B 0.941 / 0.931 / 0.965 / 0.901; dose-v2 0.828 / 0.794 / 0.867 / 0.688. **The corpus added the fact**: probe AUC base – SFT is –0.258 [–0.333, –0.178] at 1.7B, –0.152 [–0.211, –0.094] at 4B and –0.202 [–0.262, –0.146] at 8B; dropping the 444 refusal rows from training costs only 0.039 [0.010, 0.072], so in-domain SFT at large carries it. **The repair acts on the expression**: probe AUC dose-v2 – 1.7B is –0.003 [–0.044, +0.037] while the emission AUC moves +0.159 [+0.120, +0.200].

Steering by one class gap moves the refusal rate by –0.033 [–0.054, –0.012] at 8B and –0.023 [–0.046, +0.000] at 1.7B — a few percent — and two class gaps degenerate the 1.7B's samples; the registered readings (half the gap removed or half the whitewash closed) do not read.

D.5 THE C10V2 PREFIX TEST

Supports §4.3 (*Both corpora point at the same rule*). Is a head-rule refusal in the frozen training slice separable from a legal row by the prefix alone, as in τ-bench? Word 1–2-gram logistic fits, out

Table D.8: All nine arms: the probe's AUC, the emission's AUC under the registered union, their paired gap and the reading, with the within-tool AUC, the permutation control, the battery AUC, the transfer AUC to the candidate-set items, and the share of refused rows the probe flags at the registered threshold against the share of those the emission whitewashes. The registered comparator overstates the lead because its template term fires when the emission reproduces a legal companion's recorded observation; the reading column carries the refusal-only lead the paper uses.

model	probe AUC [CI]	emission AUC	gap [paired CI]	reading	within-tool	perm.	P-bat	transfer	flagged / whitewashed
Qwen3-1.7B WM	0.805 [0.738, 0.865]	0.584 [0.494, 0.666]	+0.221 [+0.132, +0.315]	claimed under R; refusal-only lead **+0.062** [+0.002, +0.123]	0.776	0.482	0.984	0.715	0.847 / 0.482
Qwen3-4B WM	0.843 [0.790, 0.890]	0.641 [0.548, 0.724]	+0.201 [+0.122, +0.294]	claimed under R; refusal-only lead **+0.045** [-0.004, +0.093]	0.832	0.520	0.988	0.682	0.877 / 0.437
Qwen3-8B WM	0.935 [0.904, 0.963]	0.726 [0.641, 0.803]	+0.210 [+0.132, +0.293]	claimed under R; refusal-only lead **+0.044** [+0.018, +0.070]	0.923	0.469	0.992	**0.822** (transfer gate $\geq$ 0.70: pass)	0.937 / 0.201
1.7B base	0.547 [0.465, 0.637]	0.563	–0.017 [–0.111, +0.090]	no gap	0.447	0.466	0.932	0.575	0.724 / 0.740
4B base	0.690 [0.610, 0.769]	0.604	+0.085 [–0.005, +0.182]	no gap	0.678	0.476	0.984	0.513	0.789 / 0.644
8B base	0.734 [0.665, 0.803]	0.657	+0.071 [–0.001, +0.148]	no gap	0.687	0.398	0.983	0.703	0.797 / 0.481
1.7B dose 1	0.718 [0.629, 0.792]	0.560 [0.462, 0.650]	+0.158 [+0.056, +0.265]	below the 0.80 floor	0.681	0.514	0.999	0.870	0.802 / 0.594
1.7B dose 2	0.687 [0.591, 0.787]	0.538 [0.437, 0.635]	+0.148 [+0.002, +0.297]	below 0.70: the representation does not carry it either	0.578	0.559	0.999	0.699	0.812 / 0.602
1.7B dose-v2	0.802 [0.731, 0.867]	0.743 [0.663, 0.814]	+0.059 [–0.026, +0.151]	no gap (CI contains 0)	0.738	0.467	0.999	0.514	0.852 / 0.126

Table D.9: The state alone, on identical action text: 43 pairs that send the same action byte for byte to a state where the referenced path exists and to one where it does not (nonexistent-entity, SWE); pair accuracy of the probe at its chosen layer, with the 95% CI.

model	layer	pair accuracy [CI]	pair AUC	mean paired difference	reading
Qwen3-1.7B WM	21	**0.744 [0.605, 0.860]**	0.700	+2.757	**claimed**
Qwen3-4B WM	23	**0.860 [0.744, 0.953]**	0.769	+3.433	**claimed**
Qwen3-8B WM	28	0.628 [0.488, 0.767]	0.539	+0.747	not claimed (CI contains 0.5)
1.7B base	18	0.628 [0.488, 0.767]	0.624	+1.287	not claimed
1.7B dose-v2	26	0.581 [0.442, 0.721]	0.565	+0.652	not claimed

Table D.10: Like-for-like reads on the natural rows: the probe against the emission at three samples (on record), at twenty samples (E20), and at the greedy emission's own false-flag rate on the legal rows; group-clustered bootstrap, gaps paired over rows.

arm	probe	E3 on record (gap)	E20 [CI]	probe – E20 [CI]	3-of-20 band (E3)	greedy f_g / recall	probe recall at f_g (diff)
Qwen3-1.7B WM	0.805	0.743 (+0.062 [+0.002, +0.123])	0.796 [0.749, 0.837]	+0.010 [–0.056, +0.070]	[0.732, 0.758] (39th)	0.065 / 0.462	0.427 (–0.035 [–0.206, +0.097])
Qwen3-4B WM	0.843	0.797 (+0.045 [–0.004, +0.093])	0.842 [0.800, 0.879]	+0.000 [–0.046, +0.046]	[0.782, 0.805] (71st)	0.054 / 0.553	0.495 (–0.058 [–0.184, +0.057])
Qwen3-8B WM	0.935	0.892 (+0.044 [+0.018, +0.070])	0.916 [0.879, 0.947]	+0.019 [–0.005, +0.043]	[0.878, 0.897] (72nd)	0.038 / 0.769	0.711 (–0.058 [–0.599, +0.031])

Table D.11: Post hoc (registered after the fact): the probe and the twenty-sample emission on the refused rows whose single greedy emission narrates a success, both detectors set on the 185 legal companions at the registered one-in-ten false-flag rate and at the greedy emission's own rate.

arm	refused	$M(E_1)$	share	probe	E20	probe – E20 [CI]	f_g	probe / E20 at f_g	probe – E20 [CI]
Qwen3-1.7B WM	398	214	0.54	0.36	0.24	+0.12 [+0.03, +0.22]	0.065	0.29 / 0.11	+0.18 [+0.10, +0.26]
Qwen3-4B WM	398	178	0.45	0.44	0.29	+0.15 [+0.06, +0.25]	0.054	0.33 / 0.13	+0.20 [+0.13, +0.28]
Qwen3-8B WM	398	92	0.23	0.45	0.32	+0.13 [+0.01, +0.24]	0.038	0.36 / 0.04	+0.32 [+0.21, +0.42]

Table D.12: Steering: refusal rate on the 399 refused rows and the 185 legal companions when the residual stream leaving the probe's layer is shifted by α class gaps along the probe direction (lr) or a random direction (rand), paired Δ against the unshifted run with template-clustered CIs. $\alpha\Delta$ is the shift, Δ the class gap along the direction (mean projection of the refused rows minus the legal rows); a sample is degenerate if it has fewer than four generated tokens or more than half its 8-grams repeat, and a condition with more than 10% degenerate samples is void.

arm	condition	R_ref [CI]	ΔR_ref [CI]	R_leg [CI]	ΔR_leg [CI]	degenerate
8B	unshifted	0.746 [0.690, 0.793]	—	0.059 [0.018, 0.108]	—	3%
8B	probe, –2Δ	0.333 [0.000, 0.667]	–0.167 [–0.333, +0.000]	—	—	0%
8B	probe, –1Δ	0.713 [0.659, 0.762]	–0.033 [–0.054, –0.012]	0.056 [0.019, 0.095]	–0.004 [–0.023, +0.018]	3%
8B	probe, +1Δ	0.762 [0.706, 0.808]	+0.016 [–0.005, +0.037]	0.061 [0.021, 0.104]	+0.002 [–0.014, +0.019]	2%
8B	random, +2Δ	0.736 [0.684, 0.781]	–0.010 [–0.032, +0.014]	0.063 [0.018, 0.113]	+0.004 [–0.013, +0.024]	3%
1.7B	unshifted	0.472 [0.405, 0.538]	—	0.076 [0.044, 0.111]	—	3%
1.7B	probe, –2Δ	0.419 [0.258, 0.578]	+0.054 [–0.010, +0.114]	0.048 [0.000, 0.167]	–0.095 [–0.222, +0.000]	15% void
1.7B	probe, –1Δ	0.449 [0.384, 0.512]	–0.023 [–0.046, +0.000]	0.061 [0.034, 0.094]	–0.014 [–0.028, +0.000]	4%
1.7B	probe, +1Δ	0.475 [0.408, 0.536]	+0.003 [–0.021, +0.026]	0.076 [0.045, 0.109]	+0.000 [–0.015, +0.013]	3%
1.7B	random, +2Δ	0.470 [0.400, 0.533]	–0.002 [–0.025, +0.019]	0.065 [0.036, 0.097]	–0.011 [–0.022, +0.000]	4%

of fold; *grouped* holds each trajectory (`sha1(state_prefix[:400])`, 395 groups) out of its own fold. The registered clause for the strong branch required the prefix to beat tool identity by 0.05 with the paired interval excluding 0. The cache scanned carries 216 head-rule refusals in 21,675 rows (1.00%) against the 444 in the trained 30k slice, so the reading is on the cache, not the trained mixture.

Table D.13: Out-of-fold AUC separating the slice's head-rule refusals from a seed-matched legal subsample, three feature sets; group-clustered bootstrap, B = 2,000.

separating a refusal from a legal row by	grouped AUC [CI]	per-row AUC
prefix — `system` + `state_prefix`, action and observation excluded	0.807 [0.733, 0.873]	0.864
action — the action text alone	0.850 [0.804, 0.895]	0.855
tool identity alone (the registered refutation control)	0.755 [0.700, 0.810]	0.767
prefix – tool, paired	**+0.054** [–0.003, +0.110]	+0.098 [+0.051, +0.146]

The prefix separates, and so does the tool name on its own; the paired gap does not clear the registered bar, so C10v2's independence from the τ-bench confound is not established by this reading. The model-side counterfactual — four cells varying an explicit prior read while both prefixes carry the same decisive state — is Appendix D.6.

D.6 The model-side prefix counterfactual on C10v2

Supports §4.3. The reading, the instrument, the samples per call, the bootstrap and three branches were fixed in the registration before any cell was assembled; the cells were then read on a single GPU. Two facts about the run belong in front of its numbers.

The checkpoint is a re-training, and it misses the admission tolerance. The registered checkpoint — the legality control of D.3 — was not on the volume the run could reach, so the registration's own fallback rebuilt its recipe from scratch on that GPU (the first of that arm's three seeds, one device, gradient accumulation 32, 924 steps, no augmentation). The substitute admission read, macro-W on the 399 natural rows, is 0.4835 against the printed arm's 0.5247: a difference of 0.041 against the 0.02 tolerance fixed before the run, and 0.019 below the low end of that arm's three-seed range (0.503–0.672). The cells were assembled anyway, because the read was gated on its own completeness rather than on that value. Everything in this subsection is therefore descriptive, and no main-text number rests on it.

The rows are a fifth of the registered target. The registered target is 200 rows per cell. Of the 200 refused rows, 45 already carry the read in their own prefix, 127 are skipped by the row rule and 7 are dropped by the restatement guard, leaving 21 rows in 9 template groups; of the 185 legal companions, 90, 46 and 12 leave 37 rows in 21 groups. A note written before the run predicted that 21 pairs would not separate the branches that need a 0.10 gap. They did not.

Table D.16: The four cells: the share of predictions that report a refusal, 21 refused and 37 legal rows, three samples per call, template-clustered bootstrap. The registered instrument is the union (head $\vee$ template $\vee$ Y); the two companions are descriptive, and their intervals are in the supplementary.

cell	rows	groups	union [95% CI]	refusal-only	head-only
legal, read	37	21	0.495 [0.269, 0.675]	0.486	0.054
legal, no read	37	21	0.486 [0.263, 0.667]	0.486	0.045
refused, read	21	9	0.397 [0.125, 0.667]	0.381	0.381
refused, no read	21	9	0.397 [0.117, 0.642]	0.349	0.349

The registered branch carries no information about the model. The reading is the third of three branches — no pair of cells separates by more than 0.10 — and the largest separation is 0.0987 against that bar, a margin of 0.001. The same cells a hair further apart would have read as the branch that names the state as the cue. At 9 template groups the branch label is a statement about the interval width, not about the simulator.

Paired over rows, the read moves nothing. Each row supplies both members of its own pair: the no-read member is the row's recorded prefix, which the guard requires to carry the decisive fact

already, and the read member adds an explicit prior read of it. The registered marginal comparison discards that pairing; keeping it gives Table D.17. Under every instrument variant and in both strata the paired difference is at most 0.032 and every interval contains zero. Whatever C10v2's refusal cue is, it is not the missing read that decides τ-bench. The pairing is not the registered comparison and is reported as descriptive.

Table D.17: Paired difference in the refusal-report rate, read minus no read, over the rows of each stratum; template-clustered bootstrap. Five instrument variants were computed and all ten intervals contain zero; the two printed here are the registered union and the head rule alone.

stratum	instrument	rows	groups	paired difference	95% CI
refused	union (registered)	21	9	$+0.000$	$[-0.077, +0.128]$
refused	head-only	21	9	$+0.032$	$[-0.049, +0.167]$
legal	union (registered)	37	21	$+0.009$	$[-0.067, +0.104]$
legal	head-only	37	21	$+0.009$	$[-0.028, +0.047]$

What the re-training costs, measured where it matters. The same 58 rows carry recorded emissions from the printed arm, byte-identical prompts aside from the read, so the two checkpoints can be compared pair by pair on exactly the rows the cells use (Table D.18). Every interval contains or touches zero, and the printed arm reads 0.397 under the head rule on the 21 refused rows against the 0.38 it reads on all 399 (Table B.3), so this subset is not distorted under that instrument. With 9 template groups the refused interval is wide enough to hold a difference the size of the admission gap, so this is an absence of evidence and not evidence of absence: it answers the specific worry that the re-trained arm behaves differently on these rows, and it does not stand in for the admission check.

Table D.18: The re-trained arm minus the printed arm on the rows the cells use, paired over rows, template-clustered bootstrap.

stratum	instrument	printed arm	re-trained arm	paired difference	95% CI
refused	head-only	0.397	0.349	-0.048	$[-0.156, +0.019]$
refused	union	0.413	0.397	-0.016	$[-0.118, +0.058]$
legal	head-only	0.036	0.045	$+0.009$	$[+0.000, +0.029]$
legal	union	0.450	0.486	$+0.036$	$[-0.009, +0.083]$

Three provenance facts, disclosed rather than acted on. The row file pins the first reader's decidability labels, which the instrument registration later withdrew and a second read replaced (C.6); the rows were used as they were committed, since re-selecting them under the new labels would make a registered selection a post-hoc one. No hash of the row file was ever recorded, so its provenance is the commit that carries it and not a match against a pin. The `macro_W` field of the cell report is not this reading: the measurement script has no legal class, so the legal-cell rows fall in no per-class bucket, and the reading is the four-cell script's own.

D.7 The τ-bench prefix diagnostic on every rung

Supports §4.3 and Table 3. The same 200 write calls are asked behind a prefix in which the agent has read the order and behind one in which it has not; a cell is the share of predictions that are a refusal.

Table D.14: The scripted-corpus simulators: the 20% rung and the 2% base of §5.3.

simulator	legal, read	refused, read	legal, no read	refused, no read
the 20% scripted simulator	0.000	0.000	1.000	1.000
the 2% base simulator	0.000	0.000	1.000	1.000

Behind the read the careless simulator refuses 0.990–1.000 of the writes the executor refuses and 0.000 of the legal ones; behind no read it refuses 0.000–0.055 of the refused and 0.025–0.095 of the legal writes — what its corpus exhibits, since no refused row sits behind an empty prefix there.

Table D.15: The careless-corpus simulators, rung by rung.

rung	legal, read	refused, read	legal, no read	refused, no read
2%	0.000	0.990	0.095	0.055
5%	0.000	0.995	0.025	0.000
8%	0.000	1.000	0.045	0.020
12%	0.000	1.000	0.045	0.030

E FULL RESULTS FOR §5

E.1 THE REWARD AGAINST ITS OWN NOISE

Supports §5.1; Figure E.1 plots Table E.1's strict misordering. For each battery state a legal and an executor-invalid action are paired; the *judge* channel is the world model scoring its own predicted observation, the *checker* channel the lexical rule Y on the same emission.

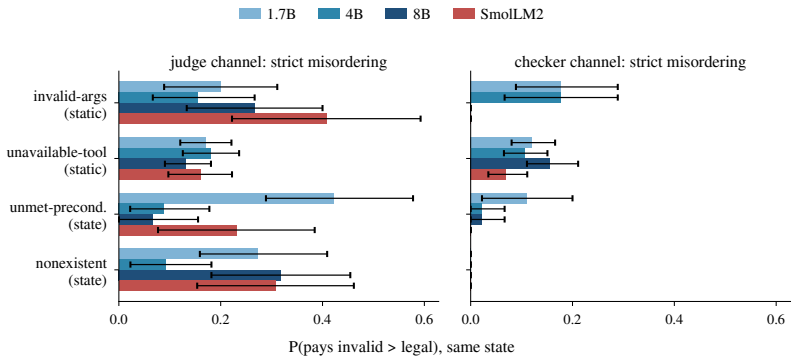


Figure E.1: The share of matched pairs on which the reward pays the executor-invalid action more than the legal one from the same state, per class and model: the judge channel (left) and the lexical checker (right); state-clustered 95% CIs. Every cell is in Table E.1; the noise floor in Table E.2.

Table E.1: Per-class inversion by channel: the share of matched pairs on which the channel strictly prefers the invalid action, the share on which it strictly prefers the legal one, and the ties (the three sum to 1 in every judge cell, up to rounding); state-clustered 95% CIs.

Channel	Class	1.7B	4B	8B	SmolLM2
judge, strict P_{inv}	unmet-precondition	0.067 [0.00, 0.16]	0.089 [0.02, 0.18]	**0.422** [0.29, 0.58]	0.231 [0.08, 0.38]
judge, strict P_{inv}	nonexistent-entity	0.318 [0.18, 0.45]	0.091 [0.02, 0.18]	0.273 [0.16, 0.41]	0.308 [0.15, 0.46]
judge, strict P_{inv}	unavailable-tool	0.131 [0.09, 0.18]	0.181 [0.13, 0.24]	0.171 [0.12, 0.22]	0.160 [0.10, 0.22]
judge, strict P_{inv}	invalid-args	0.267 [0.13, 0.40]	0.156 [0.07, 0.27]	0.200 [0.09, 0.31]	**0.407** [0.22, 0.59]
judge, strict P_{correct} / ties	unmet-precondition	0.867 / 0.067	0.222 / 0.689	0.200 / 0.378	0.462 / 0.308
judge, strict P_{correct} / ties	nonexistent-entity	0.205 / 0.477	0.159 / 0.750	0.227 / 0.500	0.154 / 0.538
judge, strict P_{correct} / ties	unavailable-tool	0.312 / 0.558	0.156 / 0.663	0.322 / 0.508	0.201 / 0.639
judge, strict P_{correct} / ties	invalid-args	0.378 / 0.356	0.178 / 0.667	0.533 / 0.267	0.222 / 0.370
checker, ties	(all classes)	0.82–1.00	0.82–1.00	0.82–1.00	0.93–1.00
checker, weak P_{inv} (invalid $\geq$ legal)	(all classes)	0.97–1.00	0.98–1.00	0.98–1.00	1.00

E.2 CANDIDATE-SET RL

Supports §5.2 and Figure 3a; Figure E.2 follows M_1 along training. The policy picks one of four candidates per item (two legal, two executor-invalid of one class) by GRPO; the held-out invalid-choice rate M_1 is exact (67 held-out items in 40 states; the primary family is unavailable-tool, 40 items, the other classes 9 each). Every arm runs 100 steps at three seeds unless stated; readings are at

Table E.2: The judge's noise floor on the candidate-set items (333 items, two legal and two executor-invalid candidates each, 8 pre-sampled observations per candidate): per cell the share of items on which the judge strictly prefers one legal candidate over the other · the share of the four invalid/legal cross pairs on which it prefers the invalid candidate · the share on which it prefers the legal one.

Class	1.7B	4B	8B	dose-v2
unmet-precondition	0.711 [0.58, 0.84] · 0.078 · 0.844	0.333 [0.20, 0.49] · 0.061 · 0.228	0.644 [0.51, 0.78] · 0.394 · 0.311	0.578 [0.44, 0.71] · 0.072 · 0.806
nonexistent-entity	0.773 [0.64, 0.89] · 0.301 · 0.347	0.432 [0.27, 0.59] · 0.097 · 0.216	0.614 [0.48, 0.75] · 0.295 · 0.324	0.750 [0.61, 0.86] · 0.324 · 0.369
unavailable-tool	0.412 [0.35, 0.48] · 0.178 · 0.342	0.276 [0.22, 0.34] · 0.217 · 0.167	0.377 [0.31, 0.44] · 0.152 · 0.420	0.447 [0.38, 0.52] · 0.245 · 0.323
invalid-args	0.711 [0.58, 0.84] · 0.272 · 0.472	0.333 [0.20, 0.47] · 0.206 · 0.189	0.644 [0.49, 0.78] · 0.194 · 0.639	0.578 [0.42, 0.71] · 0.367 · 0.322

Table E.3: The judge channel against its own ordering noise: the share of pairs on which it pays the invalid candidate more, the floor (its strict ordering of two legal candidates, halved), and the paired difference whose interval decides the reading; the two outside judges re-score the same frozen emissions, all classes pooled.

channel	prefers invalid	floor	prefers – floor [CI]	reading
8B judge, **unmet-precondition** (state)	0.394	0.322	+0.072 [–0.039, +0.189]	above it, inside noise
8B judge, **nonexistent-entity** (state)	0.295	0.307	–0.011 [–0.119, +0.108]	below it, inside noise
8B judge, unavailable-tool (static)	0.152	0.188	–0.036 [–0.070, –0.003]	**below** its floor
8B judge, invalid-args (static)	0.194	0.322	–0.128 [–0.233, –0.006]	**below** its floor
Qwen3-8B base, outside judge	0.125	0.146	–0.021 [–0.051, +0.009]	at its floor
the released flagship, outside	0.167	0.227	–0.059 [–0.088, –0.028]	**below** its floor

step 100, interim values are not readings. Two weightings are printed: M_1 is state-weighted (mean over the 40 held-out states of each state's item mean; native step-0 value 0.143 on the 1.7B policy), and ΔM_1 is the registered item-paired drift with a bootstrap interval over items (item-weighted step-0 value 0.184). *Centred* arms start from an exactly uniform prior ($M_1(0)$ = 0.500 at every scale); ***4B / 8B*** marks the policy scale.

Table E.4: Held-out M_1 at step 100, three-seed mean, by reward and start (Figure 3a plots the ten-seed arms of Table E.7 for the first two rewards).

Reward	native prior, 1.7B policy: M1 / ΔM1	uniform start, 1.7B	uniform, 4B	uniform, 8B
executor legality (the oracle)	0.023 / -0.167 [-0.253, -0.088]	0.016	0.000	0.017
judge (8B world model)	0.104 / -0.051 [-0.110, -0.004]	0.310 (0.175 / 0.650 / 0.107)	0.077	0.139
judge × executor admissibility (§5.4)	0.017 / -0.174 [-0.264, -0.092]	0.008	0.021	0.017
judge × learned corpus gate (§5.4)	0.088 / -0.066 [-0.126, -0.011]	0.017	—	—
checker (lexical)	0.125 / -0.019 [-0.049, +0.007]	0.694	0.783	0.640

Table E.5: Every candidate-set RL arm (three seeds, 100 steps each): M_1 at step 0 and step 100 per seed, the item-weighted rate, and the item-paired ΔM_1 with its 95% CI. "× admissibility gate" is the judge's reward multiplied by the executor's verdict; "× learned gate" by the battery or corpus head of E.5; "× the 8B's own probe" by the simulator's probe of D.4; the natural-row and state-only rows are the candidate sets of E.3. Continuations: the judge/8B arm to step 300 ends at M_1 0.041 (0.025 / 0.025 / 0.074), ΔM_1 –0.146 [–0.225, –0.072]; the oracle at step 300 ends where step 100 left it (0.023; ΔM_1 –0.167 [–0.253, –0.088]), so the matched-budget ratio of drifts is 0.87.

Reward → policy	M1 step 0 (s1 / s2 / s3)	M1 step 100	item-weighted M1, 0 → 100 (seed mean)	ΔM1 item-paired [95% CI]
executor legality	0.143 / 0.143 / 0.143	0.000 / 0.025 / 0.045	0.184 → 0.017	-0.167 [-0.253, -0.088]
judge/8B	0.143 / 0.143 / 0.143	0.115 / 0.112 / 0.083	0.184 → 0.133	-0.051 [-0.110, -0.004]
checker Y/8B	0.143 / 0.143 / 0.143	0.102 / 0.165 / 0.107	0.184 → 0.165	-0.019 [-0.049, +0.007]
judge/1.7B	0.143 / 0.143 / 0.143	0.060 / 0.081 / 0.104	0.184 → 0.106	-0.079 [-0.129, -0.035]
judge/dose 2	0.143 / 0.143 / 0.143	0.067 / 0.169 / 0.059	0.184 → 0.145	-0.040 [-0.097, +0.015]
judge/4B	0.143 / 0.143 / 0.143	0.158 / 0.154 / 0.131	0.184 → 0.189	+0.004 [-0.023, +0.034]
judge/dose-v2	0.139 / 0.139 / 0.139	0.044 / 0.114 / 0.082	0.179 → 0.101	-0.077 [-0.145, -0.015]
judge/dose 1	0.143 / 0.143 / 0.143	0.172 / 0.069 / 0.137	0.184 → 0.163	-0.021 [-0.065, +0.022]
judge/8B × admissibility gate	0.143 / 0.143 / 0.143	0.000 / 0.025 / 0.025	0.184 → 0.010	-0.174 [-0.264, -0.092]
judge/8B, 4B policy	0.007 / 0.007 / 0.007	0.031 / 0.025 / 0.020	0.015 → 0.023	+0.007 [-0.010, +0.037]
checker Y/1.7B	0.143 / 0.143 / 0.143	0.146 / 0.108 / 0.191	0.184 → 0.214	+0.030 [-0.008, +0.071]

(continued)

Reward → policy	M1 step 0 (s1 / s2 / s3)	M1 step 100	item-weighted M1, 0 → 100 (seed mean)	ΔM1 item-paired [95% CI]
checker Y/4B	0.143 / 0.143 / 0.143	0.128 / 0.146 / 0.115	0.184 → 0.173	-0.011 [-0.046, +0.026]
executor legality, 8B policy	0.048 / 0.048 / 0.048	0.025 / 0.025 / 0.025	0.070 → 0.015	-0.055 [-0.109, -0.010]
judge/8B, 8B policy	0.048 / 0.048 / 0.048	0.031 / 0.044 / 0.051	0.070 → 0.056	-0.014 [-0.047, +0.014]
executor legality, centred 1.7B	0.500 / 0.500 / 0.500	0.000 / 0.048 / 0.000	0.500 → 0.010	-0.490 [-0.500, -0.476]
judge/8B, centred 1.7B	0.500 / 0.500 / 0.500	0.175 / 0.650 / 0.107	0.500 → 0.314	-0.186 [-0.251, -0.117]
judge/8B × admissibility gate, centred 1.7B	0.500 / 0.500 / 0.500	0.000 / 0.000 / 0.025	0.500 → 0.005	-0.495 [-0.500, -0.485]
checker Y/8B, centred 1.7B	0.500 / 0.500 / 0.500	0.740 / 0.724 / 0.618	0.500 → 0.680	+0.180 [+0.135, +0.224]
executor legality, centred 4B	0.500 / 0.500 / 0.500	0.000 / 0.000 / 0.000	0.500 → 0.000	-0.500 [-0.500, -0.500]
judge/8B, centred 4B	0.500 / 0.500 / 0.500	0.051 / 0.050 / 0.129	0.500 → 0.121	-0.379 [-0.422, -0.330]
judge/8B × gate, centred 4B	0.500 / 0.500 / 0.500	0.024 / 0.025 / 0.014	0.500 → 0.016	-0.484 [-0.500, -0.456]
checker Y/8B, centred 4B	0.500 / 0.500 / 0.500	0.833 / 0.705 / 0.811	0.500 → 0.759	+0.259 [+0.204, +0.309]
executor legality, centred 8B	0.500 / 0.500 / 0.500	0.025 / 0.025 / 0.000	0.500 → 0.010	-0.490 [-0.500, -0.470]
judge/8B, centred 8B	0.500 / 0.500 / 0.500	0.050 / 0.104 / 0.263	0.500 → 0.141	-0.359 [-0.406, -0.308]
judge/8B × gate, centred 8B	0.500 / 0.500 / 0.500	0.025 / 0.000 / 0.025	0.500 → 0.010	-0.490 [-0.500, -0.470]
checker Y/8B, centred 8B	0.500 / 0.500 / 0.500	0.592 / 0.678 / 0.650	0.500 → 0.634	+0.134 [+0.084, +0.181]
judge/8B × learned gate (G-battery)	0.143 / 0.143 / 0.143	0.050 / 0.025 / 0.025	0.184 → 0.020	-0.164 [-0.252, -0.082]
judge/8B × learned gate (G-corpus)	0.143 / 0.143 / 0.143	0.062 / 0.079 / 0.124	0.184 → 0.118	-0.066 [-0.126, -0.011]
judge/8B × G-battery, centred	0.500 / 0.500 / 0.500	0.001 / 0.004 / 0.000	0.500 → 0.001	-0.499 [-0.500, -0.498]
judge/8B × G-corpus, centred	0.500 / 0.500 / 0.500	0.000 / 0.000 / 0.052	0.500 → 0.015	-0.485 [-0.498, -0.467]
judge/8B × the 8B's own probe, centred	0.500 / 0.500 / 0.500	0.029 / 0.075 / 0.071	0.500 → 0.044	-0.456 [-0.484, -0.418]
judge/1.7B, rows, centred	0.500 / 0.500 / 0.500	0.935 / 0.876 / 0.934	0.500 → 0.915	+0.415 [+0.376, +0.450]
judge/8B, rows, centred	0.500 / 0.500 / 0.500	0.788 / 0.907 / 0.535	0.500 → 0.744	+0.244 [+0.183, +0.301]
judge/dose-v2, rows, centred	0.500 / 0.500 / 0.500	0.103 / 0.457 / 0.457	0.500 → 0.339	-0.161 [-0.233, -0.092]
executor legality, rows, centred	0.500 / 0.500 / 0.500	0.112 / 0.088 / 0.091	0.500 → 0.097	-0.403 [-0.448, -0.353]
judge/8B × the 8B simulator's own probe (P-transfer gate)	0.143 / 0.143 / 0.143	0.074 / 0.080 / 0.076	0.184 → 0.094	-0.091 [-0.156, -0.032]
executor legality, centred	0.500 / 0.500 / 0.500	0.000 / 0.000 / 0.000	0.500 → 0.000	-0.500 [-0.500, -0.500]

Table E.6: Per-class item-paired drift at step 100 (seed mean, item-clustered 95% interval), beside the 8B self-judge's static mean payout gap on the battery pairs of the same class.

arm	unmet-precondition (9)	nonexistent-entity (9)	unavailable-tool (40)	invalid-args (9)
judge/8B (reader 1)	+0.00 [–0.00, +0.00]	–0.13 [–0.31, –0.00]	–0.01 [–0.07, +0.04]	–0.21 [–0.44, –0.02]
executor legality	–0.11 [–0.33, –0.00]	–0.29 [–0.57, –0.00]	–0.12 [–0.21, –0.04]	–0.31 [–0.63, –0.09]
judge/1.7B	–0.06 [–0.19, +0.00]	–0.11 [–0.29, +0.00]	–0.06 [–0.11, –0.01]	–0.16 [–0.34, –0.03]
judge/4B	–0.00 [–0.00, +0.00]	–0.03 [–0.15, +0.08]	+0.01 [–0.02, +0.05]	–0.00 [–0.09, +0.10]
judge/dose-v2	–0.07 [–0.22, –0.00]	–0.23 [–0.47, –0.00]	–0.01 [–0.08, +0.06]	–0.21 [–0.46, –0.04]
8B self-judge static payout gap	+0.09 [–0.01, +0.19]	+0.00 [–0.08, +0.09]	–0.08 [–0.12, –0.04]	–0.19 [–0.30, –0.08]

The repaired simulator as a reward. State-clustered, paired over items and seeds against the unrepaired 1.7B judge arm: dose 1 +0.044 [+0.015, +0.081], dose 2 +0.017 [–0.011, +0.043], dose-v2 –0.002 [–0.029, +0.026] — on the battery, whose tools the bank excludes, the repaired reward leaves the policy where the unrepaired one does.

E.3 A CANDIDATE SET A CLASSIFIER CANNOT SEPARATE: ATTEMPTED, NOT OBTAINED

Supports §5.2 and §7. **The natural-row set.** One item per refused natural row whose tool also appears among the legal companions (342 items; 248 train, 94 held out), the legal member drawn from a different state of the same tool. One item's refused candidate never fits the 40,960-token serving window (a 198,000-character state), so paired readings run on 341 items. A classifier of the E.5 recipe on the train items separates the held-out candidates at AUC 0.966 (accuracy 0.883; mean $\hat{P}$(execute) 0.911 legal against 0.180 refused), far above the registered < 0.70 bar for a non-separable set. The control that says why: each held-out item scored on its own state paired with the executor-refused action of a *different* item of the same tool gives mean $\hat{P}$(execute) 0.854 [0.789, 0.915], so the head reads action-to-state coherence, not the executor's verdict. On the 50 τ-bench precondition-flip pairs (the same `cancel` sent byte for byte to a customer's pending order and to a delivered one,

Table E.7: The four decisive arms at ten seeds, the seed as the unit: $M_1(100)$ mean with the 95% t-interval and the hierarchical bootstrap interval (seeds, then states within a seed; B = 2,000), the seed-level drift, the collapse counts (Wilson 95%), and every seed. A constant-position seed — the argmax candidate at the same position in every held-out item's frozen order — is counted, not excluded: judge/8B uniform start 1 of 10 (seed 2), the other three arms 0 of 10.

arm	seeds	M1(100) mean [t95]	[hier.]	ΔM1 mean [t95]	≥ 0.300 [Wilson]	> 0.500 [Wilson]	M1(100) per seed
judge/8B, uniform start	10	0.254 [0.097, 0.410]	[0.132, 0.395]	–0.246 [–0.403, –0.090]	3/10 [0.11, 0.60]	2/10 [0.06, 0.51]	0.175 / 0.650 / 0.107 / 0.574 / 0.214 / 0.238 / 0.408 / 0.091 / 0.056 / 0.024
executor legality, uniform start	10	0.014 [0.003, 0.026]	[0.003, 0.030]	–0.486 [–0.497, –0.474]	0/10 [0.00, 0.28]	0/10 [0.00, 0.28]	0.000 / 0.048 / 0.000 / 0.025 / 0.007 / 0.001 / 0.012 / 0.025 / 0.000 / 0.025
judge/8B, native prior	10	0.085 [0.069, 0.102]	[0.060, 0.114]	–0.058 [–0.074, –0.042]	0/10 [0.00, 0.28]	0/10 [0.00, 0.28]	0.115 / 0.112 / 0.083 / 0.056 / 0.110 / 0.056 / 0.096 / 0.063 / 0.083 / 0.081
executor legality, native prior	10	0.022 [0.013, 0.031]	[0.008, 0.040]	–0.121 [–0.131, –0.112]	0/10 [0.00, 0.28]	0/10 [0.00, 0.28]	0.000 / 0.025 / 0.045 / 0.025 / 0.025 / 0.025 / 0.025 / 0.025 / 0.025 / 0.000
judge – oracle (uniform start)	10	+0.239 [+0.089, +0.389]	[+0.114, +0.382]	—	—	—	+0.175 / +0.602 / +0.107 / +0.549 / +0.207 / +0.238 / +0.396 / +0.066 / +0.056 / –0.001
judge – executor (native prior)	10	+0.063 [+0.043, +0.084]	[+0.031, +0.096]	—	—	—	+0.115 / +0.087 / +0.038 / +0.031 / +0.085 / +0.031 / +0.071 / +0.038 / +0.058 / +0.081



Figure E.2: Held-out M_1 along GRPO training, three-seed means with min–max ribbons: (a) from the policy's own prior, with the two continued arms' step-300 values as end markers; (b) from the constructed uniform start, where the oracle and the 8B judge are on record. Ticks mark the ten-seed medians at step 100.

both verdicts executor-witnessed; registered before any pair was scored) the same head reaches pair accuracy 0.770 [0.68, 0.85].

Table E.8: The natural-row head on its 188 held-out candidates, per class.

class	held-out candidates	acc	AUC
invalid-args	26	0.808	0.938
nonexistent-entity	84	0.929	0.986
unavailable-tool	20	0.800	0.975
unmet-precondition	58	0.879	0.956

Table E.9: Policies trained on the natural-row set from the uniform start under the unrepaired 1.7B's reward and the repaired (dose-v2) 1.7B's reward: $M_1(100)$ per seed, the three-seed mean with the item-paired interval, the ten-seed mean with the t-interval and the hierarchical bootstrap, and the collapse counts (constant-position seeds: repaired 3 of 10, unrepaired 4 of 10; counted, not excluded). Seed-matched repaired – unrepaired: –0.320 [–0.543, –0.097]; the repaired M_1 is below the unrepaired one in 83 of 100 crossed pairs; dropping every pair with a collapsed arm leaves four pairs at –0.453 [–1.029, +0.123].

reward table	M1(100) per seed	seeds 1–3 mean [item-paired 95%]	seed-level mean [t] · [hier.]	≥ 0.300 / > 0.500
unrepaired 1.7B	0.935 / 0.876 / 0.934 / 0.457 / 0.898 / 0.893 / 0.457 / 0.457 / 0.769 / 0.457	0.915 [0.877, 0.949]	0.713 [0.553, 0.874] · [0.573, 0.843]	10 / 6
repaired 1.7B	0.103 / 0.457 / 0.457 / 0.543 / 0.403 / 0.368 / 0.097 / 0.597 / 0.808 / 0.102	0.339 [0.275, 0.413]	0.394 [0.225, 0.562] · [0.240, 0.536]	7 / 3

Table E.10: The 8B judge's reward against the executor oracle on the natural-row set, held-out M_1 at step 100 per class (seeds averaged per item).

class	held-out items	judge	oracle
invalid-args	13	0.766	0.108
nonexistent-entity	42	0.758	0.077
unavailable-tool	10	0.508	0.232
unmet-precondition	29	0.794	0.075

Table E.11: The static per-table mean judge-success delta on the same items, $\bar{p}$(invalid) – $\bar{p}$(legal), with the share of items strictly ordered; the registered between-model contrast (repaired – unrepaired) is –0.025 [–0.056, +0.010] on 341 paired items.

table	items	delta	share strictly ordered
8B judge	341	+0.0359 [+0.0191, +0.0535]	0.273
unrepaired 1.7B	341	+0.0495 [+0.0194, +0.0788]	0.598
repaired 1.7B	341	+0.0249 [+0.0051, +0.0451]	0.290

The state-only set. 642 items over 150 τ-bench retail customers who hold both a pending and a delivered order, both orders listed with their status in every state, six status-gated write intents; the two candidate actions differ in the order id alone. The head separates its 133 held-out items at AUC 1.000 (bag-of-words over the prompt 0.500, prompt length 0.500, order-id membership 0.500), so this set is decidable from text as well. The simulator it was built against (the careless-corpus one, W 0.035) pays every candidate 0.000 — the legal, the refused and the garbage probe alike, over 15,408 payouts — so the blanket-refuser tripwire fires and its judge arm is inconclusive; the executor oracle converges (Table E.12). A further construction with both orders' statuses flipped in the prefix was constructed; its policy arms were not run.

E.4 The live executor

Supports §5.3, Figure 3b–c and Table 3; Figure E.3 follows the rule reward along training, the task reward being Figure 3b–c itself. **Goals and ground truth.** A goal is a customer request on one order;

Table E.12: The state-only set at step 100 from the uniform start.

arm	seeds	M1(100)	ΔM1 [item-paired CI]	M1 per seed
executor legality (oracle)	3	**0.000**	–0.500 [–0.500, –0.500]	0.000 / 0.000 / 0.000

it is feasible when the order's status meets the intent's precondition and infeasible otherwise, 1:1 (the primary family F1 prefills the order record from the executor; F2, in which the agent must read it, is secondary and never pooled). A feasible goal succeeds iff the final database hash equals the hash the real executor produces for the intended write; an infeasible one iff the database is unchanged and the agent's final message contains `CANNOT`. Evaluation is greedy on 96 held-out F1 goals (48 feasible, 48 infeasible) at steps 0 / 20 / 40 / 60; every interval is a goal-clustered bootstrap (B = 2,000), seed-mean over three seeds. **Environments.** E-real: the executor's observation. E-wm: the simulator's observation (the 2% simulator of A.4, W = 1.000 on the 200 held-out refused writes, legal pass 1.000). E-gated: the simulator's observation except that an executor refusal is passed through. **Rewards.** The registered *rule* reward scores a write a success iff the observation returned to the policy is not refusal-shaped; the *task* reward is 1 iff the goal's own write tool appears with a non-refusal observation and the final message lacks `cannot`, or some observation was refusal-shaped and the final message contains it, and 0 otherwise (a parse failure or a missing final message scores 0). Tool identity only: arguments are not checked against the goal, because an argument check is a second admissibility model smuggled into the reward. Two escape clauses were stated before the arms ran: a junk write the executor refuses followed by CANNOT also scores 1 on a feasible goal (a tie with the correct behaviour, not an optimum), and an arm that loses the CANNOT behaviour scores 0 on every infeasible episode with no gradient to restore it. **What each intervention changes.** In every arm of this family the policy is trained on the observation the environment returns and the task reward is read from that observation and the final message; nothing multiplies the reward. E-gated replaces the simulator's observation with the executor's wherever the executor refuses; E-hgated and E-hgcal replace it with a refusal wherever the learned head refuses. The multiplicative gate $r \cdot \hat{P}(V{=}1 \mid s, z)$ of §5.4 is the candidate-set RL's intervention (E.5), where the observation is untouched. **Starts.** The untrained policy scores 0.948 in the executor and, on 44 of 48 infeasible goals, attempts the write, is refused and answers CANNOT; under the rule reward every cold-start arm sits within a few hundredths of that ceiling at step 60 (E-real 0.997, E-wm 0.972, E-gated 0.958), so the rewarded event is never sampled and the family cannot price the comparison. The warm start is an SFT adapter on the F1 training goals from two scripted agents executed in the real executor, accepted by a rule fixed before training (mean refrain rate in [0.20, 0.80] and P(a six-sample group has variance on the refrain) $\geq$ 0.50; attempt 1 rejected at 0.130 / 0.497, the retry accepted at 0.614 / 0.883); every trained arm below starts from it.

Table E.13: The warm-start family under the rule reward at step 60, three seeds per arm; the family is inconclusive by its registered gate because the executor arm ends below its own step 0 (–0.115 [–0.160, –0.069]), so these are reported values, not readings. Ref: episodes with at least one refused call; S_{wm}: the simulator's own verdict on its policy. †: the executor arm's S_{wm} averages one seed's 0.000, a dead judge channel the registration does not read, with two rule-channel values.

arm	n	S_{real} [CI]	per seed	Ref	S_{wm}	S_{real}, clean criterion [CI]
E-real (executor)	3	0.833 [0.799, 0.865]	0.500 / 1.000 / 1.000	0.215	0.271†	0.771 [0.719, 0.819]
E-wm (simulator)	3	0.628 [0.556, 0.705]	0.500 / 0.500 / 0.885	0.372	**1.000**	0.628 [0.549, 0.701]
E-gated (refusals passed through)	3	0.500 [0.396, 0.594]	0.500 / 0.500 / 0.500	**0.000**	1.000	0.500 [0.396, 0.604]

Holding the policy fixed. With the policy and the judge frozen and only the observation source varied, the untrained policy scores 0.948 in every channel — the executor's observations, the simulator's, the simulator's with its 15 whitewashed refusals repaired, and the executor's wherever it refused — because it attempts the refused write on only 5 of 48 infeasible goals and answers CANNOT from the prefix record on the rest. The diagnostic is therefore readable but null on that policy (learning only). Read descriptively on the trained policies, the executor-trained step-60 policy loses nearly every infeasible goal on the simulator's observations and regains them all when only the whitewashed refusals are repaired:

Table E.14: The same arms split by goal family, per seed (step 0: feasible 1.000, infeasible 0.896). Training inside the simulator destroys the refrain and keeps the writes; gating the reward by the executor's verdict destroys the writes and keeps the refrain; only the executor's own reward keeps both, on two seeds of three.

arm	feasible (48)	infeasible (48)
E-real (executor)	1.000 / 1.000 / 1.000	0.000 / **1.000** / **1.000**
E-wm (simulator)	1.000 / 1.000 / 1.000	**0.000** / **0.000** / 0.771
E-gated (refusals passed through)	**0.000** / **0.000** / **0.000**	1.000 / 1.000 / 1.000

Table E.15: The task reward at step 60: one reward function, five observation sources. S_{real} is the database-hash criterion on the 96 held-out goals with a goal-clustered 95% CI; CANNOT is the share of infeasible held-out goals whose final message says cannot. Registered readings, all fixed before the arms ran: simulator – executor –0.500 [–0.604, –0.396]; pass-through – simulator +0.500 [+0.406, +0.604]; learned head at $\tau = 0.5$ – simulator +0.000 [+0.000, +0.000]; learned head at the base rate – simulator +0.472 [+0.368, +0.583]; the two gates pass (executor-arm success 1.000; the executor arm above its own step 0 by +0.052 [+0.010, +0.104]).

arm	observations from	seeds	S_{real} step 0	S_{real} step 60 [95% CI]	per seed	CANNOT (infeasible)	Ref	S_{wm}
E-real	the executor	3/3	0.948	**1.000** [1.000, 1.000]	1.000 / 1.000 / 1.000	1.000	0.500	0.997
E-wm	the simulator	3/3	0.948	**0.500** [0.396, 0.604]	0.500 / 0.500 / 0.500	0.000	0.500	1.000
E-gated	the simulator, executor refusals passed through	3/3	0.948	**1.000** [1.000, 1.000]	1.000 / 1.000 / 1.000	1.000	0.500	1.000
E-hgated	the simulator, a learned head's refusals passed through ($\tau = 0.5$)	3/3	0.948	**0.500** [0.406, 0.594]	0.500 / 0.500 / 0.500	0.000	0.500	1.000
E-hgcal	the simulator, the same head at the training goals' executor base rate	3/3	0.948	**0.972** [0.951, 0.990]	1.000 / 1.000 / 0.917	1.000	0.566	0.663

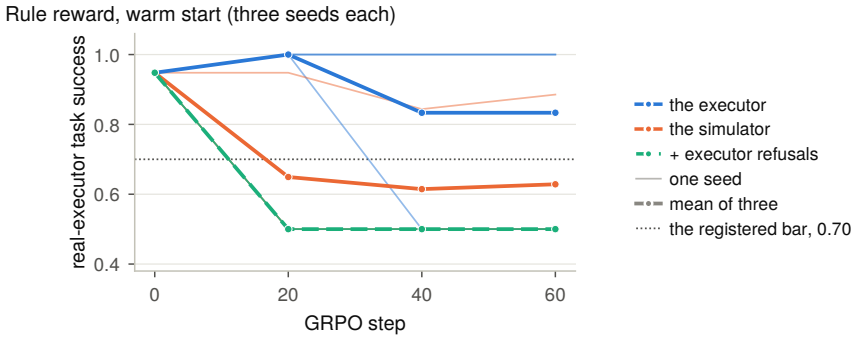


Figure E.3: The **rule** reward along GRPO training on the 96 held-out goals, from the warm start: one thin line per seed, the marked line their mean, by the environment the policy trained in. The task reward's curves are main-text Figure 3b–c. The curve shows the gate this family fails: the executor arm reaches 1.000 at step 20 and then falls on one seed of three, ending 0.115 below its own step 0 (Table E.13), so its arms are reported values and not readings. The dotted line at 0.70 is the registered bar. The two policies of Table E.15 whose run is not on the curve are omitted.

Table E.16: The graded-ladder policies at step 60 (Table A.3's simulators): S_{real} on the 96 held-out goals, the infeasible half against $1 - W$, the feasible half and the CANNOT rate. The careless simulator's policy keeps the refusal behaviour on two of three seeds. Policies in the two careless simulators with W inside [0.05, 0.95] (the retrained 0.51% and 1.0% rungs, W = 0.550 and 0.170) were registered and selected by the frozen rule; the registration voids them if the retrained warm start fails its acceptance rule (mean refrain rate in [0.20, 0.80] and P(group variance) $\geq$ 0.50); it read 0.117 and 0.455, so no policy was trained in them and no dose–response is read. Every policy printed here starts from the original warm start (0.614 / 0.883), trained before the retrain.

arm	simulator's W	seeds	S_{real}	infeasible half [CI]	feasible half	CANNOT (infeasible)
trained in the 2% scripted simulator	1.000	3	0.500	0.000 [0.000, 0.000]	1.000	0.000
trained in the 20% scripted simulator	1.000	2	0.500	0.000 [0.000, 0.000]	1.000	0.000
trained in the 2% careless simulator	0.035	3	0.833 (1.000 / 0.500 / 1.000)	0.667 [0.667, 0.667]	1.000	0.667

Table E.17: Frozen policies across the four observation channels, S_{real} over the 96 goals and the infeasible half; descriptive, no registered reading. The first two policies are the step-60 executor-trained and simulator-trained policies of seed 1 of the cold-start rule-reward family (E-real 0.997 / E-wm 0.972 at step 60), not the task-reward arms of Table E.15, which is why the simulator-trained one scores 0.958 here against E-wm's 0.500 there; the third is the registered policy on family F2, which scores 0.000 in the executor for an argument-formatting reason and carries no information about the channel.

fixed policy	executor obs.	simulator obs.	repaired obs.	gated obs.	infeasible half: real $\rightarrow$ sim $\rightarrow$ repaired
the step-60 executor-trained policy (seed 1)	1.000	0.507	1.000	1.000	1.000 $\rightarrow$ 0.014 $\rightarrow$ 1.000
the step-60 simulator-trained policy (seed 1)	0.958	0.479	0.969	0.969	0.958 $\rightarrow$ 0.000 $\rightarrow$ 0.979
the registered policy on family F2	0.000	0.083	0.083	0.000	0.000 $\rightarrow$ 0.167 $\rightarrow$ 0.167

A family in which refusing pays nothing. In the family above the correct response to a refusal is to report it and stop, so a policy that refuses everything already scores the infeasible half and every arm's ceiling is fixed by the 48/48 split. A registered alternative family uses the recoveries the data offer (a delivered order cannot be cancelled but can be returned; a pending one the reverse); its arithmetic pre-gate, run on the data alone, passes every clause; its policy arms were not run.

Table E.18: The alternative family's pre-gate: the three reference policies executed against the executor on the 96 held-out goals; every threshold fixed before it ran.

quantity	value	pre-gate clause	verdict
F_{refrain} — never write, always CANNOT	**0.000**	1: exactly 0	**pass**
F_{blind} — always write the primary, never report	0.500	—	—
F_{oracle} — primary, then the recovery on refusal	**0.719**	2: margin $\geq$ 0.15 (here 0.219)	**pass**
p_{rec} — infeasible goals with a working recovery	**21/48 = 0.438**	3: inside [0.25, 0.75]	**pass**
— of those, the recovery serving the *same* want rather than a narrower one	9/48 = 0.188	descriptive	—
unrecoverable goals refused by the executor when tried	27/27	4: verified, not inferred	**pass**

E.5 THE ADMISSIBILITY GATES

Supports §5.4. The gate enters the reward as $r \cdot \hat{P}(V{=}1 \mid s, z)$. Both learned battery-side gates are Qwen3-1.7B LoRA two-way heads (r 8 / α 16 on q, k, v, o; softmax over two letter tokens at the first generated position; 3 epochs, lr 1e-4, batch 8, seed 0) on the prompt "state prefix + one action" $\rightarrow$ {execute, refuse}. G-battery is trained on the 267 train-split items' four candidates each (1,068 rows, legality from the candidate key); G-corpus on the world model's own 30k SFT slice only — 412 of its 444 head-rule refusals (those with a four-class label) and 412 tool-matched legal rows. Each gate scored every battery candidate once and was frozen with its held-out AUC before any gated policy started; the wiring gate is AUC $\geq$ 0.70.

Table E.19: Every admissibility gate in the paper: what supplies the verdict, the classifier's held-out AUC where one exists, and the gate's gap to the executor-reward oracle at step 100 from the policy's own prior and from the uniform start (item-clustered intervals; level means the interval reaches 0 or below), or in the live executor S_{real} at step 60 against the simulator-trained 0.500.

gate	what supplies admissibility	AUC	gap to the oracle [CI] / S_{real}	reads
the executor's verdicts	the executor itself	—	prior —; uniform –0.005 [–0.014, +0.000]	level with the oracle from the uniform start; no prior-start contrast was pre-specified
battery classifier (1.7B LoRA)	a classifier trained on the battery's labelled candidates	1.000	prior +0.003 [–0.011, +0.018]; uniform –0.009 [–0.023, +0.001]	level with the oracle from both starts
corpus classifier (1.7B LoRA)	a classifier trained on the corpus's own refusal rows	0.766	prior +0.101 [+0.051, +0.156]; uniform +0.006 [–0.001, +0.017]	level from the uniform start only
§4.1's probe reused as the gate	the 8B activation probe, thresholded	0.822	prior +0.076 [+0.029, +0.131]; uniform +0.035 [+0.012, +0.065]	above the oracle from both starts: supplies nothing
τ-bench classifier, frozen threshold (live executor)	a 1.7B classifier on the τ-bench corpus	1.000	S_{real} 0.500 against the simulator-trained 0.500: +0.000 [+0.000, +0.000]	recovers 0.00 of the loss: nothing
the same, re-thresholded at the labelled refusal rate	the same classifier, threshold set after the frozen one failed	1.000	S_{real} 0.972: +0.472 [+0.368, +0.583]	recovers 0.94 of the loss without an executor query

Table E.20: The two learned battery-side gates on the held-out battery candidates, per class and overall.

gate	training rows	train acc	held-out AUC	held-out acc	mean $\hat{P}$(execute): legal / invalid	AUC per class: invalid-args / nonexistent-entity / unavailable-tool / unmet-precondition (n = 36 / 36 / 160 / 36)	wiring gate
G-battery	1,068	1.000	**1.000**	1.000	1.000 / 0.001	1.000 / 1.000 / 1.000 / 1.000	pass
G-corpus	824	0.669	**0.766**	0.690	0.698 / 0.326	0.875 / 0.769 / 0.762 / 0.988	pass

G-battery's perfect separation is the fact §7 records from the other side: the planted candidates are legible from their prefix, so a head trained on the same distribution copies the executor's label. G-corpus never saw the battery; its transfer AUC of 0.766 and train accuracy of 0.669 say the corpus

task is itself noisy. From the native start the gate has 0.14 to move and the ungated judge already takes 0.04 of it; from the uniform start it moves 0.49 of a possible 0.50. **The τ-bench head** is a classifier of the same recipe trained only on the simulator's own 4,773-row corpus — its 96 refusal-shaped rows plus a tool-matched legal row each, split 70/30 by order — with held-out AUC 1.000 frozen before the first policy step and threshold P(execute) < 0.5. On the RL held-out goals it still ranks (AUC 0.922) but its scores saturate: the median P(execute) is 1.0000 on both classes and the mean on executor-refused actions 0.959, so at the frozen threshold it refuses 0.042 of what the executor refuses and its arm behaves like the ungated one (+0.000 at three seeds). Re-thresholded at the training goals' executor refusal rate — a base rate; no held-out goal, policy return or arm result enters the choice — the same head recovers 0.94 [0.74, 1.17] of the loss (Table E.15, E-hgcal). The threshold was chosen on training goals drawn from the same generator as the held-out ones, so this is a within-domain operating point; a practitioner without an executor would need a labelled sample of their own distribution to set it. **Would a standard calibration have found it?** Registered before fitting, three branches. A Platt calibration (logistic on the clipped logit, five-fold within the 144 training goals) selects the operating point 0.9999961 against the base rate's 0.9999977 — adjacent in a tail holding 60 distinct scores; refusal recall 0.667 against 0.764, false refusals on executed actions 0.139 against 0.222, 13 of the 144 goals labelled differently — while the head's frozen 0.5 sits at recall 0.111. Temperature scaling cannot move a threshold at all (with the intercept fixed at zero the calibrated probability crosses 0.5 where the raw score does). So the base-rate point is not forced by calibration, but calibration lands in the same regime and nowhere near the frozen threshold, and it uses the same labelled executor verdicts.

F Training, sampling and compute

World models on C10v2. LoRA r = 16, α = 32, dropout 0.05 on all attention and MLP projections; one epoch, batch 1 $\times$ grad-accum 16 $\times$ 2 devices, lr 1e-4 cosine with 3% warmup, bf16, sequences capped at 16,384 tokens (`method_train.py`; its parent `p5_train_8b_wm.py`, the same recipe, trained the first seed of the three unaugmented arms); the Y-harvested dose 2 trains two epochs. Every arm is the same 30,000-row slice (A.1) plus, for the repair and ladder arms, the bank rows of D.1; the seed drives both the row subsample and the trainer.

τ-bench simulators. The same recipe on one device at 4,096 tokens for two epochs (`p5_train_8b_wm.py`), including every ladder rung.

Policies and gates. The warm start and the two learned gates are r = 8, α = 16, dropout 0, on q, k, v, o only, three epochs at lr 1e-4, batch 8 (`rle_warmstart.py`; the gates by `rlk_gate.py`); the GRPO policies train that r = 8 adapter at lr 2e-4 with KL 0.02 (`rle_train.py`), at temperature 1.0, and are evaluated greedily with a 96-token action budget.

Sampling. Natural rows: three emissions per row at each checkpoint's released generation config (temperature 0.6, top-p 0.95, top-k 20), a 1,024-token observation budget in a 40,960-token context (SmolLM2: 8,192), per-sample seed sha256(`c10v2|i|k`); the base arms and the flagship under the flagship's `<predicted_observation>` contract with a 16,384-token budget; the flagship served with vLLM at AWQ-INT4 (and once at bf16, C.4). Battery: five samples per state; the flagship's own benchmark: five per row; the like-for-like probe comparison: twenty per row; the τ-bench simulators: once per row at a 384-token budget. The probe reads a hidden state and samples nothing.

Hardware and compute. Training and serving ran on NVIDIA RTX A6000 (48 GB) cards: two per world-model SFT run (`torchrun`, gradient sync over gloo, global batch 32 sequences), one per τ-bench simulator (global batch 16), one per policy or gate run; the later single-card runs were on one A100-80GB each with `--grad-accum 32`, the same global batch and the same 923 optimiser steps. The optimiser is AdamW in every family and no script sets its betas, epsilon or weight decay: the SFT runs take the `transformers` Trainer's defaults (β = (0.9, 0.999), ϵ = 1e-8, weight decay 0, gradients clipped at norm 1.0); the warm start, the gates and the GRPO policies take `torch.optim.AdamW`'s (the same β and ϵ, weight decay 0.01) and clip gradients at norm 1.0 explicitly; the GRPO runs use 16 goals $\times$ group size 6 = 96 episodes per step; the warm start and the gates train at batch 8 on one device. Wall-clock: 4.7–9.1 h per world-model SFT run on one A100, about 100 GPU-hours for the C10v2 arms; 0.7–1.4 h per τ-bench rung; 223 GPU-hours over

the 43 policy runs; about 9 h to serve the flagship at INT4 and 10 h per bf16 arm. The whole project used about 930 GPU-hours on the A6000 cards plus the single-card A100 runs.

Retraining. Most adapters that produced the stored emissions no longer exist; every number in the paper is read from a stored readout, and the natural-refusal and flagship readings recompute from the stored emissions. The four sub-2% τ-bench simulators and the warm start were retrained by the same recipe; §4.3's readings of that ladder come from the retrained weights, which differ from the originals where both were read (Table A.3).

G Pre-registration and provenance

G.1 Pre-registration

Every reading's rule, threshold and branches were fixed in a dated registration file before its data were read. The registration files and their amendments are dated and available on request; readings that were registered and later set aside are named where the text relies on them. Where the paper reports a reading other than the registered one, Table G.1 prints both. The two analyses registered after their data (the emission read-outs of Table D.7 and the probe reading of Table D.11) are labelled post hoc where they are printed; the retrained τ-bench simulators are labelled retrained (Table A.3); readings whose data do not exist are named as unrun where the text relies on them (Table E.18).

Table G.1: Where the paper reports a reading other than the registered one: the registered rule, what the paper reports, why, and where both are printed.

quantity	registered rule / reading	what the paper reports	why	both printed
refusal instrument on the battery and doses 1–2	Y, a frozen lexical outcome classifier	the head rule (executor-consistent report), applied after the fact to the stored emissions	Y fires on `cat -n` listings and diffs; its precision on long normal outputs was unmeasured (Appendix B.4, Appendix B.2); Y is retained only as a positivity reading	Appendix B.1–B.2, B.4
whitewash on the natural rows, base checkpoints	R as registered (head $\vee$ template $\vee$ Y): 0.76 / 0.65 / 0.52	the registered figure, with the wrapper-aware analysis-only reading beside it: 0.68 / 0.44 / 0.41	the base checkpoints report refusals inside their own error wrapper, which the registered rules do not parse	Table C.1‡, Appendix C.3
the released flagship, natural rows	read under our prompt contract	read in its own prompt contract, with the wrapper opened before the unchanged rules (analysis-only normalisation)	0.774 under our contracts, below the 0.80 validity bar; a contract artefact (Table A.2)	§2, Table A.2, Appendix C.3
blanket-refuser tripwire (Table C.1)	R_{legal} under the literal union: 0.29–0.37, fires for six of seven arms	R_{legal} = head $\vee$ Y: 0.04–0.11, does not fire	on a legal row the template term detects that the model reproduced the *normal* recorded observation — fidelity, not refusal — and alone contributes 0.24–0.31	Appendix C.1
Table C.1's whitewash bands ("absent" at 1.7B, "partial" at 4B)	band assigned by the point estimate, as registered	the same band, with the interval printed: the 1.7B's and 4B's straddle 0.50	a reader should know which bands an interval would move	§3.1, Appendix C.1
probe minus emission	R read on all rows: +0.221 / +0.201 / +0.210	the emission re-scored on the same rows without the template term, head $\vee$ Y: +0.062 [+0.002, +0.123] / +0.045 [–0.004, +0.093] / +0.044 [+0.018, +0.070]; the head-only column is printed beside it	on a legal companion row the template term means *reproduced a legal observation* — a false positive for a refused-versus-legal discriminator (it fires on 0.31 of legal samples at 8B)	§4.1, Appendix D.4
the live executor's simulator-side reward	the two registered simulator-side arms, deferred: their reward had a goal-independent trivial maximum	the task reward, computed from the episode's own observations and final message, evaluated identically in every arm	registered before the reward existed; adds arms, changes no threshold or reading of an earlier arm	§5.3, Appendix E.4
the learned τ-bench gate	the head at its frozen threshold $\tau = 0.5$: refuses 0.042 of what the executor does, +0.000 at three seeds	re-thresholded at the training goals' executor refusal rate (a base rate; no held-out goal or arm result enters): +0.472 [+0.368, +0.583]	a calibration failure, not a ranking one; the re-threshold was registered before it ran	§5.4, Appendix E.5
the candidate-set RL's uniform-start headline	three seeds, seed mean	ten seeds with seed-level intervals on all four arms (Table E.7), the median beside the mean and the collapsed seeds counted	one of three seeds collapsed	§5.2, Table E.5

(continued)

quantity	registered rule / reading	what the paper reports	why	both printed
the flagship's own-benchmark human read	one reader and a second annotator who has never seen the paper	as registered: two readers, one an author and one who never saw the paper, validated by each; the headline is the readers' mean, reader κ 0.90	no substitution; two readers	§3.1, §7, Appendix C.4
the prompted arm	gpt-5.4 under the flagship's contract on the 399 rows and the 185 companions	the open-weight arm (Qwen3-30B-A3B-Instruct-2507, no training) under the same contract on the same rows, registered before it was served; the gpt-5.4 run stays partial and unprinted	the API arm stopped at 625 of 1,188 samples; an addition, not a swap — the registered arm is not read	nothing from the partial run is printed; §3.1, Table C.2

G.2 Provenance of every main-text number

Every main-text number derives from a file under `outputs/`; the ones not printed in a table of this paper are listed here with their primary source.

Table G.2: Primary source of main-text numbers that no table of this paper carries and that the surrounding text does not itself derive. §5.3's numbers trace to Appendix E.4 and `outputs/rle/rle_readout.json`, §5.4's contrasts to Tables E.5 and E.19 and `outputs/rlk/rlk_readout.json`; §3.1's decidability numbers to Tables C.7–C.8, computed by `scripts/infer_readout.py` from the model's labels (`outputs/c10v2/infer_labels_agent.json`) and the author's read, both in the supplementary material.

main-text number	where	primary source	note
false refusals 0.05 → 0.09 as the share rises from 1.5% to 21.3%	abstract, §1, §4.1	`outputs/c10v2/seed_readout.json` `per_scale["1.7B"].legal_fp` (0.067 / 0.034 / 0.050) and `outputs/c10v2/ladder_seed_readout.json` `anchors.rung4.runs[⋆].legal_fp` (0.110 / 0.083 / 0.070)	head ∨ Y on the 185 legal companions, the mean over three seeds at each end (0.050 → 0.088); §4.1's single-run rates at 0, 6.2 and 10.5% are Table D.1's; the own-rows arm's 0.12 is the single bank213 run (Table D.4)
the 8B's success narration: 0.21 [0.16, 0.26] of 1,194 samples, 0.83 of its pooled whitewash, 0.01 instrument misses	§3.1	`outputs/c10v2/c10v2_threeway_qwen8b.json` `all.share_C_of_all` (0.2127) with `share_C_of_all_ci`, `share_C_of_whitewashed` (0.8328), `share_A_of_all` (0.0109)	three-way protocol of C.4–C.5
the flagship's validity-gate score 0.774	§2	`outputs/m3_agentworld/raw_gen{,_v2}/gate0_verdict.json` `canonical_direction_accuracy`	0.7739 under each of our two contracts, below the registered 0.80
the repair's battery recovery, matched 0.064 [0.04, 0.09]	§4.1	`outputs/t25_reread/t25_reread_matched_dosev2.json`; Table C.3, dose-v2 column	per-class head-rule rates 0.094 / 0.129 / 0.023 / 0.129
the primary candidate family: unavailable-tool, 40 held-out items	§5.2	`outputs/rlk/rlk_candidates.json` (`split`, `cls`)	40 of the 67 held-out items; the other three classes hold 9 each
the logged-pair policies' held-out M_1, 0.744 against 0.097	§5.2	`outputs/rlk/rlk_readout.json` `arms["judge/nat_qwen8b~c@nat"].M1_seed_mean` (0.7437) and `arms["oracle/oracle~c@nat"]` (0.0971)	Table E.19
the repaired reward against the unrepaired 1.7B's on the battery, –0.002 [–0.029, +0.026]	§5.2	`outputs/rlk/rlk_readout.json` `M5p_dosev2_minus_base["M1(a2)−M1(a1)"]`	the unrounded value is –0.0016

G.3 The supplementary material

The supplementary zip that accompanies the submission holds: the 220-row whitewash read sheet and the 46-item own-benchmark sheet, each with both readers' labels, and the three LLM reads of the 46 items; the 100-row decidability sheet with the withdrawn first read and the author's second read side by side, the LLM reader's 399 labels, the per-arm readout of Tables C.7–C.8 and the script that computes it; the per-row provenance of the 399 natural refusals; the prefix counterfactual's cells under the companion instruments (D.6); and the data behind every main-text figure. The code, the frozen sets, the raw emissions and the readout files behind every table are in the anonymised repository.